Biological sequence analysis by vector maps: alignment-free comparison of DNA and proteins

Susana Vinga Martins

Dissertation presented to obtain the Doutoramento (Ph.D.) degree in Biology
at the Instituto de Tecnologia Química e Biológica of
Universidade Nova de Lisboa

Oeiras, February 2005

Apoio financeiro da FCT e do FSE no âmbito do
Quadro Comunitário de apoio, BD nº3134/2000

Biological sequence analysis by vector maps:
alignment-free comparison of DNA and proteins

ISBN 972-99615-0-6

PhD Thesis Public Discussion in Biology
Susana Vinga Martins
25th February 2005, at 14:30 H

"Biological sequence analysis by vector maps: alignment-free comparison of DNA and proteins"

Presidente (President of the Jury)
Doutor Peter Frank Lindley, Presidente do Conselho Científico do ITQB

Vogais (Members of the Jury):
– Doutora Marie-France Sagot, Directeur de Recherche, Institut National de Recherche en Informatique et en Automatique (INRIA), France
– Doutor Gregory Warr, Professor Catedrático, Medical University of South Carolina, USA
– Doutor Jonas Silva Almeida, Professor Associado, Medical University of South Carolina, USA
– Doutor Carlos Daniel Mimoso Paulino, Professor Associado com Agregação do Instituto Superior Técnico da Universidade Técnica de Lisboa
– Doutor Mário Nuno Ramos d'Almeida Ramirez, Professor Auxiliar da Faculdade de Medicina da Universidade de Lisboa
– Doutora Ana Teresa Correia de Freitas, Professora Auxiliar do Instituto Superior Técnico da Universidade Técnica de Lisboa
– Doutor Jorge Albino Cadeias Araújo Carneiro, Investigador Auxiliar do Instituto de Tecnologia Química e Biológica da Universidade Nova de Lisboa

From left to right: Peter Lindley, Jonas Almeida, Mário Ramirez, Susana Vinga, Daniel Paulino, Gregory Warr, Jorge Carneiro, Marie-France Sagot, Ana Teresa Freitas. *Oeiras, 25th February 2005.*

Things are easier in practice than in theory
João Janeira

To Amarildo, Lina,
Inês and Simone

Acknowledgments

I would like to thank my supervisor, Prof. Jonas Almeida, for transmitting his constant optimism in science and for his encouragement and support throughout these years.

I thank all the Biomathematics Group for the excellent environment I had and all the fun and work we managed to do nonetheless: Francisco Pinto, João Carriço, João Xavier, Rodrigo Gouveia-Oliveira, Sara Garcia, Sara Silva. I gratefully acknowledge the work of António Maretzek as system administrator of our group, who managed all the computers used in this thesis. I also thank Isabel Bahia for additional computer support and Ana Maria Portocarrero and Cristina Amaral for so efficiently organizing the Thesis defense and presentation.

I would like to thank Prof. Marie-France Sagot for greeting me in her group in Lyon and so enjoyably sharing her research projects. I also thank Prof. Ana Teresa Freitas for her availability and welcoming me in her projects.

I would like to thank Prof. Daniel Paulino for helpful comments during the preparation of one paper and Prof. Jorge Carneiro for giving useful suggestions during a presentation in his group.

I acknowledge the financial support of Fundação para a Ciência e a Tecnologia (FCT), which made this work possible (grant SFRH/BD/3134/2000).

I would like to thank the people in Charleston who so nicely welcomed me: Kim Pawlick and Kathleen Russell for sharing their houses and also Angela, Noé and Raquel Almeida, Renee Hutson and Greg Vick.

I thank André, Bart, Clara, Filipe, Jota, Júnior, Maria João, Maria José, Martin, Nheca, Ricardo, Sara, Teresa and Tiago for listening to my (sometimes boring) scientific discussions and *queixinhas*. I thank Mimi for visiting me abroad. I also thank Xé-xé for his feline presence during all my student years.

And finally I thank Amarildo, Lina, Inês and Simone, for things this page is too small to contain.

Abstract

Biological sequence analysis is one of the main bioinformatics sub-disciplines, bringing together several fields, from computer science to probability and statistics. Its purpose is to computationally process and decode the information stored in biological macromolecules involved in all cell mechanisms of living organisms – such as DNA and proteins – and provide prediction tools to reveal their structure, function and complex relationship networks.

This thesis addresses sequence analysis by vector maps, which are functions that transform sequences onto n-dimensional vectors in $\mathbb{R}^n$. These techniques do not depend on sequence alignment algorithms, which are ubiquitously used in bioinformatics applications, such as the BLAST procedure. The vector maps considered define a category, named "alignment-free", that although less explored in the literature, constitutes an important subject with significant contributions in the past years, given their natural formulation, elegant formalism and low computational cost.

Two types of functions are exploited in this work: the first one maps sequences onto their sub-string or L-tuple frequency vectors and the second one, chaos game representation (CGR), is anchored on iterative function systems (IFS) and fractal geometry theory, mapping symbols onto points with applicable topological and stochastic properties.

Following a bibliographic review of alignment-free methods, an extensive quantitative analysis of these word-composition distances is performed, along with the introduction of a new dissimilarity measure between proteins. The W-metric bridges alignment metrics and those based solely in L-tuple composition, by combining, in quadratic forms, aminoacid composition and mutational information given by substitution matrices. The evaluation of the dissimilarity measures previously reviewed is applied to the recognition of protein relationships specified by the SCOP database, a benchmark for protein hierarchical secondary structure classification.

In the study of CGR maps, the method is first extended to accommodate higher-length alphabets, named Universal Sequence Map (USM), allowing the representation of proteins and natural languages texts. CGR/USM generalizes any order Markov chain transition probability tables and is related to binary representation of numbers. In addition it holds noteworthy context properties, with suffixes far apart in the original sequence mapped onto contiguous regions and the ability of recovering all the sequence from just one point. They constitute the foundation of a new entropy measure of DNA sequences here presented. The Rényi continuous entropy of DNA sequences is based on CGR/USM and in non-parametric kernel density estimation with Parzen's window method. This entropy measure is tested on artificial and real DNA and its asymptotical behavior is deduced, along with Monte Carlo simulations performed to estimate the variability of this quantity. All the computer code described was developed in MATLAB$^{\text{TM}}$language and is made available online.

This work helps systematize alignment-free techniques by presenting an extensive review of these methods and applications, with a strong emphasis on uniform nomenclature and formalism that will support future developments in

this area. Additionally, a full quantitative analysis of dissimilarity measures obtained through these vector maps showed that although less sensitive and specific than alignment algorithms, they perform reasonably well which, associated with their extremely low computational cost, make them potentially important for data pre-filtering or heuristics improvement. A precise protocol for classification accuracy assessment was established which might be used to study other dissimilarity measures in the future. The vector maps (USM) generalized in this work motivated a novel measure of sequence entropy, which is in agreement with information theory and simulation studies and allows the study of uncertainty and predictability of biological sequences. It might be further applied to the computation of sequence entropic profiles and convey useful local information for prediction and classification problems.

The thesis, based on published papers, is organized in the following structure: *Chapter 1 – Introduction* – presents background information on molecular biology, sequence analysis and mathematical and computational methods used, such as information theory, vector maps, iterative function systems (IFS) and chaos game representation (CGR).

The following *Chapter 2 – Alignment-free sequence comparison – a review* – constitutes a bibliographic review of the main techniques for measuring sequence dissimilarity not requiring their pre-alignment. Moreover, it provides additional background information on words in sequences and strengthens the motivation for all the subsequent work. In *Chapter 3 – Universal sequence map (USM) of arbitrary discrete sequences* – a natural extension of CGR maps is identified, allowing the representation of higher-order alphabet sequences. The representation for backward sequences is explored and a dissimilarity measure between symbol mappings is proposed.

The next two chapters are devoted to applications of these methods to biological sequences. The work presented in *Chapter 4 – Comparative evaluation of word composition distances for the recognition of SCOP relationships* – refers to the quantitative assessment of classification accuracy of the dissimilarity measures previously reviewed. It also proposes a new word composition measure, the W-metric, which bridges alignment-free and alignment-based concepts. *Chapter 5 – Rényi continuous entropy of DNA sequences* – presents a CGR/USM-driven entropy definition, based on Rényi formalism, which constitutes a novel application of iterative maps for measure the uncertainty of DNA.

Chapter 6 – Final discussion – finalizes by bringing together the conclusions of previous chapters and summarizing the main contributions of this work for the analysis of biological sequences. This closing chapter also describes open problems and future developments in this area.

This report presents and expands on work described in the following publications: Vinga, S. & Almeida, J. (2003) *Bioinformatics* 19, 513–523; Almeida, J. S. & Vinga, S. (2002) *BMC Bioinformatics* 3, 6; Vinga, S., Gouveia-Oliveira, R. & Almeida, J. S. (2004) *Bioinformatics* 20, 206–215; Vinga, S. & Almeida, J. S. (2004) *J. Theor. Biol.* 231, 377–388.

Análise de sequências biológicas por funções vectoriais: comparação sem alinhamento de ADN e proteínas

Resumo

A análise de sequências biológicas é uma das áreas mais importantes da bioinformática que combina diversos campos científicos, desde as ciências da computação à probabilidade e estatística. Tem como objectivo o processamento computacional e a descodificação da informação armazenada nas macromoléculas biológicas, tais como o ADN e as proteínas, envolvidas nos mecanismos celulares de todos os seres vivos e, também, a criação de ferramentas para a predição da sua estrutura, função e inferência das complexas redes de interacção entre essas mesmas moléculas.

Esta tese propõe uma abordagem à análise de sequências por funções vectoriais que transformam o espaço das sequências em vectores n-dimensionais de $\mathbb{R}^n$. Estas técnicas não dependem de algoritmos de alinhamento, usados extensivamente em aplicações bioinformáticas, e.g. no programa BLAST. Estas funções definem uma categoria, denominada 'sem alinhamento' (alignment-free) que, embora menos explorada na literatura, constitui uma área com inúmeras aplicações importantes nos últimos anos, pela sua formulação natural, formalismo elegante e custo computacional reduzido.

Neste trabalho são explorados dois tipos de funções: a primeira transforma sequências nos seus vectores de composição, ou seja, nas frequências de ocorrência das palavras de tamanho L (L-tuples); a segunda função, representação por jogos de caos (chaos game representation – CGR), baseia-se em sistemas de funções iterativas (iterated function systems – IFS) e em geometria fractal, transformando símbolos em pontos com propriedades topológicas e estocásticas relevantes aplicáveis ao estudo da sequência original.

Após uma revisão bibliográfica de métodos sem alinhamento, é apresentada uma análise quantitativa dessas métricas – baseadas em composição de palavras – com a introdução de uma nova medida de dissemelhança entre proteínas. A métrica-W (W-metric) combina métodos com e sem alinhamento através da utilização de formas quadráticas de frequência de aminoácidos associadas a matrizes de substituição com informação evolutiva. A avaliação das medidas de dissemelhança anteriormente revistas é aplicada para o reconhecimento de relações entre proteínas especificadas pela SCOP, uma base de dados de referência para a classificação hierárquica da estrutura secundária de proteínas.

No estudo de mapas CGR, este método é inicialmente generalizado de forma a acomodar alfabetos com maior cardinalidade através de mapas de sequências universais (universal sequence maps – USM), permitindo, deste modo, a representação de proteínas e de textos em linguagem natural. CGR/USM generalizam tabelas de transição de cadeias de Markov de qualquer ordem, estão relacionadas com a representação binária de números e possuem propriedades de contexto importantes; por exemplo, os sufixos, mesmo se separados na sequência original, são aplicados em regiões contíguas, sendo também possível recuperar

toda a sequência a partir de apenas uma única coordenada. Este método constitui os fundamentos de uma nova medida de entropia de sequências, também apresentada. A entropia contínua de Rényi de sequências ADN é baseada em mapas CGR/USM e na estimação não paramétrica de densidades pelo método das janelas de Parzen. Esta medida de entropia é testada em sequências de ADN artificiais e reais e é deduzido o seu comportamento assimptótico. São efectuadas, também, simulações Monte Carlo, com o intuito de estimar a variabilidade desta medida. Todos os algoritmos descritos foram implementados em MATLAB$^{\mathrm{TM}}$e estão disponíveis online.

Este trabalho permite sistematizar o estudo de técnicas sem alinhamento, ao apresentar uma revisão extensiva destes métodos e da sua respectiva aplicação, com especial ênfase dado às uniformizações da nomenclatura e formalismo que irão auxiliar o desenvolvimento futuro desta área. Adicionalmente, a análise quantitativa exaustiva desses mesmos métodos e respectivas medidas de dissemelhança obtidas através das funções vectoriais a eles associadas, comprovam que, embora com menos sensibilidade e especificidade do que algoritmos baseados em alinhamento, se obtêm resultados com custo computacional reduzido, o que os torna potencialmente importantes para pré-processamento ou filtragem de sequências e melhoria de heurísticas existentes. Foi, também, estabelecido um protocolo para avaliação dos classificadores que poderá ser aplicado facilmente no futuro ao estudo de outras medidas de dissemelhança. A representação USM, generalizada neste trabalho, motivou a criação de uma nova medida de entropia de sequências que revelou estar concordante quer com a teoria de informação, quer com estudos de simulação, permitindo o estudo da incerteza e previsibilidade de sequências biológicas. Poderá ser aplicada, no futuro, ao cálculo de perfis entrópicos e fornecer informação local para problemas de previsão e classificação.

Esta tese é baseada em artigos publicados e tem a seguinte estrutura: o Capítulo 1 – *Introduction* – apresenta uma breve introdução à biologia molecular, análise de sequências e a diversos métodos matemáticos e computacionais, tais como teoria da informação, funções vectoriais, sistemas de funções iterativas (IFS) e jogos de caos (CGR).

O Capítulo 2 – *Alignment-free sequence comparison – a review* – é uma revisão bibliográfica das principais medidas de dissemelhança que não requerem técnicas de alinhamento. Adicionalmente, apresenta material suplementar teórico em sequências, fortalecendo, também, a motivação geral deste trabalho. No Capítulo 3 – *Universal sequence map (USM) of arbitrary discrete sequences* – é proposta uma extensão natural dos mapas CGR permitindo, assim, a representação de sequências com alfabetos de maior dimensão. É desenvolvida a representação da sequência invertida e propõe-se, também, uma medida de distância entre as imagens de símbolos.

Os capítulos seguintes apresentam aplicações destes métodos ao estudo de sequências biológicas. O trabalho apresentado no Capítulo 4 – *Comparative evaluation of word composition distances for the recognition of SCOP relationships* – refere-se à avaliação quantitativa da precisão dos classificadores baseados nas mediadas de dissemelhança revistas anteriormente. Também é proposta uma nova medida, a métrica-W (W-metric), que combina conceitos de algorit-

mos com e sem alinhamento. O Capítulo 5 – *Rényi continuous entropy of DNA sequences* – apresenta uma medida de entropia baseada em mapas CGR/USM e no formalismo de Rényi, constituindo uma nova aplicação de mapas iterativos para o estudo da incerteza de sequências de ADN.

O Capítulo 6 – *Final discussion* – conjuga as conclusões dos capítulos anteriores e recapitula as principais realizações deste trabalho para o estudo de sequências biológicas. Este capítulo final também descreve alguns problemas em aberto nesta área e algumas previsões acerca do seu desenvolvimento futuro.

Esta tese apresenta e desenvolve trabalho descrito nas seguintes publicações: Vinga, S. & Almeida, J. (2003) *Bioinformatics* 19, 513–523; Almeida, J. S. & Vinga, S. (2002) *BMC Bioinformatics* 3, 6; Vinga, S., Gouveia-Oliveira, R. & Almeida, J. S. (2004) *Bioinformatics* 20, 206–215; Vinga, S. & Almeida, J. S. (2004) *J. Theor. Biol.* 231, 377–388.

Preface

The present thesis describes work developed in bioinformatics – a recently emerging discipline at the interface of biology, informatics and statistics – presented at Instituto de Tecnologia Química e Biológica (ITQB) to obtain a PhD degree in biology.

A brief preliminary note is warranted regarding the format of this report. This thesis is a compilation of articles already published in peer-reviewed scientific journals, which is a mandatory condition to every ITQB PhD candidate. Therefore two options could have been followed related to the structure and presentation of the published material; either integrally transcribe the papers, excluding any additional data, or, alternatively, provide a more general contextualization of the work, with the inclusion of supplementary material, e.g. annexes, appendixes and an extended introduction. The first option would fulfill the minimum requirements for ITQB PhD candidates but would make for a compilation of multidisciplinary work without the benefit of an extended and less specialized presentation of its context. We have decided to follow the second solution, even at the risk of misrepresenting original material as the recontextualization itself did not enjoy the advantages of peer-review. Moreover, we think this format has the additional advantage of exploring and broadening issues not covered in the original papers due to editorial restrictions. Therefore, this preface serves to note that the original reports are always referred to as the sources for any material presented in this report.

The rationale for the approach followed is that, being of multidisciplinary nature, this thesis risked including elements that pose to experts in either area the inconvenience of frequently seeking external introductory material. Accordingly, we have included a general introduction contextualizing the main topics covered in an attempt to produce a self contained report, instead of simply presenting the collection of papers that in fact make the substance of the thesis. We are also aware that we may be incurring on the obvious drawback of boring the biologists with the biology and the mathematicians with mathematics. Therefore, experts in either field are well advised to skip the introductory material to their own fields, included with a didactic intent for non specialists and serve, in the future, as consultation material to new scientists in the area.

Susana Vinga Martins
December 2004

Contents

Acknowledgments vii
Abstract ix
Resumo xi
Preface xv

1 Introduction **1**
1.1 Bioinformatics in the post-genomic era 2
1.2 Molecular biology and genetics 3
1.2.1 Biological sequences 3
1.2.2 Synthesis of macromolecules and the central dogma . . . 8
1.3 Sequence analysis and comparison 11
1.3.1 Alignment methods for sequence comparison 11
1.3.2 Vector maps for sequence comparison 14
1.4 Iterative Function Systems 15
1.4.1 Definitions 15
1.4.2 Chaos game representation (CGR) 17
1.5 Entropy and information theory 22
1.5.1 Shannon's entropy function 23
1.5.2 Rényi's entropy generalization 25
1.6 Thesis outline 27
1.7 References 29

2 Alignment-free sequence comparison – a review **33**
2.1 Introduction 34
2.2 Background 35
2.2.1 Words in sequences 36
2.2.2 Distance between sequences 36
2.2.3 Word statistics 37
2.2.4 Information theory 37
2.3 Alignment-free sequence comparison 38
2.3.1 Methods based on word frequencies 39
2.3.2 Resolution free methods 45
2.4 Algorithm implementation –NASC-Toolbox 47
2.4.1 Matlab functions 48
2.4.2 DNA 49
2.4.3 Proteins 49
2.4.4 Natural languages 54

2.5 Conclusions . . . 58
2.6 Acknowledgements . . . 58
2.7 References . . . 58

3 Universal sequence map (USM) 65
3.1 Background . . . 66
3.2 Results . . . 66
3.2.1 Conceptual foundations . . . 67
3.2.2 Implementation of USM algorithm . . . 67
3.3 Discussion . . . 75
3.4 Conclusions . . . 79
3.5 Methods . . . 79
3.5.1 Computation . . . 79
3.5.2 Source code . . . 79
3.5.3 Test data . . . 79
3.6 Acknowledgements . . . 80
3.7 References . . . 80

4 Comparative evaluation of word composition distances 83
4.1 Introduction . . . 84
4.2 Systems and Methods . . . 85
4.2.1 Word statistics . . . 85
4.2.2 W-metric definition . . . 86
4.2.3 ROC curve definition . . . 87
4.2.4 Protein test datasets – SCOP/ASTRAL classification . . . 88
4.2.5 Protocol for comparative assessment . . . 89
4.2.6 Computation . . . 90
4.3 Results and Discussion . . . 90
4.3.1 Complete dataset . . . 90
4.3.2 Stratified analysis by class . . . 95
4.4 Conclusion . . . 97
4.5 Acknowledgements . . . 98
4.6 References . . . 98

5 Rényi continuous entropy of DNA sequences 101
5.1 Introduction . . . 101
5.2 System and methods . . . 104
5.2.1 CGR/USM representation of a sequence . . . 104
5.2.2 Rényi continuous entropy definition . . . 106
5.2.3 Parzen window density estimation . . . 107
5.2.4 Simplification of Rényi entropy calculation for USM maps 107
5.2.5 Asymptotic properties of H_2 and random sequence simulation . . . 108
5.2.6 DNA sequence dataset description . . . 109
5.3 Results and Discussion . . . 110
5.3.1 Rényi continuous quadratic entropies H_2 . . . 111
5.3.2 Equivalent sequence length N_{eq} . . . 114

5.3.3 Comparison between continuous and discrete measures of entropy . . . 115
5.3.4 Algorithm implementation – Rényi-Toolbox . . . 116
5.4 Conclusions . . . 116
5.5 Acknowledgements . . . 117
5.6 Appendix . . . 117
A. Gaussian or Normal distribution function definition . . . 117
B. Rényi quadratic entropy simplification . . . 118
C. Asymptote calculation . . . 118
D. Convolution of Normal distribution functions . . . 120
5.7 References . . . 123

6 Final discussion **127**
6.1 References . . . 130

Index **133**

List of Tables

1.1 DNA/RNA extended alphabet 4
1.2 Aminoacid alphabet 6
1.3 The genetic code 9

2.1 NASC toolbox MATLAB files 48
2.2 DNA Sequences from E. coli K12 threonine operon 50
2.3 Human beta globin sequences 51
2.4 EU languages 56
2.4 EU languages (cont.) 57

3.1 Binary codes for USM units in the two stanzas 68

4.1 Protein datasets for dissimilarity measures evaluation 89

5.1 Sequence DNA dataset used for entropy estimation 110
5.2 Rényi toolbox MATLAB files 117

List of Figures

1.1 Growth of GenBank DNA database . . . 2
1.2 DNA structure and composition . . . 5
1.3 Protein hierarchical structure . . . 7
1.4 Macromolecule synthesis in eukaryote cells . . . 8
1.5 Protein synthesis and RNA roles in translation . . . 10
1.6 Pairwise alignment example . . . 13
1.7 The Sierpinski triangle or gasket . . . 16
1.8 CGR of human beta globin region (HUMHBB) . . . 19
1.9 CGR and Markov chains . . . 22
1.10 Rényi entropy of a two state probability distribution . . . 26
1.11 Thesis outline scheme . . . 28

2.1 Dendrogram of E. coli threonine operon . . . 49
2.2 Human beta globin genes (HUMHBB) on chromosome 11 . . . 51
2.3 GenBank entry of HUMHBB . . . 52
2.4 Dendrogram of HUMHBB . . . 53
2.5 Ancestry of the hemoglobin genes . . . 53
2.6 Multiple alignment of HUMHBB proteins . . . 54
2.7 Dendrogram with EU languages classification . . . 55

3.1 USM representation of two stanzas . . . 70
3.2 Cross-tabulation of similarity between positions of the two stanzas 71
3.3 Probability distribution of similarity estimates . . . 74
3.4 Cumulative distribution of bi-directional similarity . . . 76
3.5 Comparison of uni and bi-directional USM implementation for DNA sequences . . . 78

4.1 SCOP/ASTRAL db – hierarchical classification of proteins . . . 88
4.2 ROC curves for PDB40-v dataset . . . 91
4.3 ROC curves for PDB40-b dataset . . . 92
4.4 AUC values for PDB40-v dataset . . . 93
4.5 AUC values for PDB40-b dataset . . . 93
4.6 Stratified analysis by class in PDB40-v dataset . . . 96
4.7 The α-helix and β-sheet content (%) in PDB40-b, grouped by SCOP class. . . . 97

5.1 Chaos Game Representation (CGR) suffix property . . . 105

5.2 Rényi entropy for the sequence DNA dataset 111
5.3 Rényi entropy for random simulated sequences 112
5.4 Derivative of Rényi entropy for the sequence DNA dataset 113
5.5 Derivative of Rényi entropy for random simulated sequences . . . 114
5.6 Quantile order or probability values of Rényi entropy for the sequence DNA dataset 115
5.7 Shannon discrete entropies for the sequence DNA dataset 116

Chapter 1

Introduction

Biological sequence analysis is at the core of bioinformatics, being its oldest sub-discipline. Although some of the initial paradigms are changing and new integrative techniques are being developed, it is thought to be true that sequence determines structure that in turn determines molecular function and the overall biological role of the cell's molecules. Interestingly, this discipline is also posing new problems and challenges to statisticians and computer scientists, with the development of new algorithms and conjectures that are directly inspired by open questions in biology.

This thesis addresses the field of biological sequence analysis by vector maps of DNA and proteins onto $\mathbb{R}^n$. These 'alignment-free' techniques were far less explored in the literature than alignment methods, but constitute an important subject with significant applications. Their natural formulation and elegant formalism, along with a wide range of potential applications and low computational cost make them suitable in many circumstances, such as to optimize other methods and to uncover different structural levels and properties of biological sequences.

This introductory chapter describes the thesis general motivation and overviews background information to contextualize and interconnect the remaining chapters. Its main goal is to make the reader acquainted with the major problems in biological sequence analysis and serve as a consultation guide for basic definitions and nomenclature issues.

The first section introduces the emerging field of bioinformatics, followed by a primer on biological sequences and their crucial importance in all living organisms processes, which fully justifies their study. Afterwards an introduction to sequence analysis is presented, along with the comparison between alignment methods and vector mappings proposed in this thesis. The alignment-free category will explore iterative functions systems and entropy measures, also reviewed. Finally, the last section of this chapter presents the thesis outline, describing its overall structure and suggested reading lines.

A note should be made about the thesis format, which is based on independent papers. Each of the subsequent chapters transcribes the material published in scientific peer-reviewed journals thus having a specific rigorous structure. Due to this reason, the overlapping of information throughout this work is inevitable

and several topics will be covered in more than one section. For this reason this background section will refer to those specific sections when those repetitions would be more evident and will be limited to the presentation of the information not addressed afterwards.

1.1 Bioinformatics in the post-genomic era

Bioinformatics has emerged as a new scientific field due to the great increase of biological data generation, particularly of genetic datasets. The recent genome sequencing projects created an enormous quantity of data and gave rise to an urgent need of new techniques and algorithms for analyzing the massive amounts of information thus produced. As an example, Fig. 1.1 shows the exponential growth of GenBank[1] nucleotide sequence database (Benson et al., 2004), an annotated collection of all publicly available DNA sequences maintained by the National Institutes of Health (NIH).

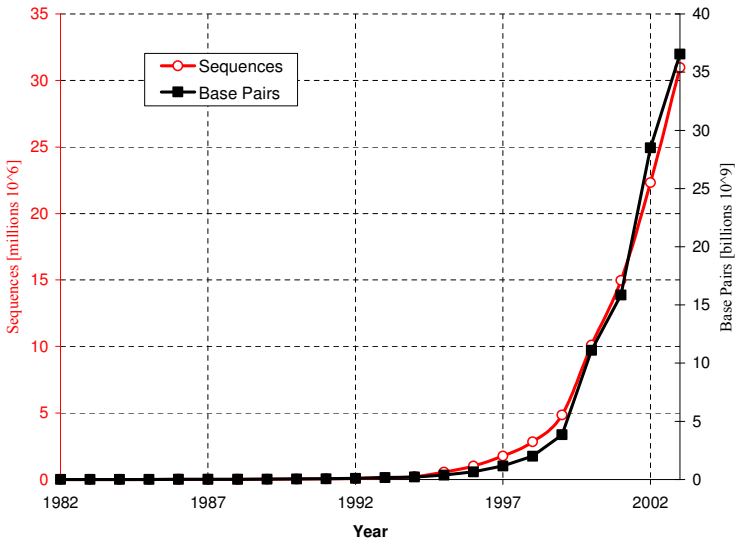


Figure 1.1: Growth of GenBank DNA sequence database.

Given the incessant discoveries and development of new techniques and algorithms, the definition of bioinformatics and computational biology is still evolving. This fact confirms its novelty and shows that bioinformatics has been widening its scope and aims, developing continuously. Nevertheless, it continues to be rooted and overlap with computer science and information technology, probability and statistics and biology. According to the definition of the NIH[2]:

> *Bioinformatics – Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.*

[1] http://www.ncbi.nlm.nih.gov/Genbank/
[2] NIH Bioinformatics Web Site - http://www.bisti.nih.gov/

> *Computational biology – the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.*

The main subdisciplines and goals of bioinformatics proposed by the National Center for Biotechnology Information[3] include: 1) the development of new algorithms and statistics with which to assess relationships among members of large data sets; 2) the analysis and interpretation of various types of data including nucleotide and aminoacid sequences, protein domains, and protein structures; 3) the development and implementation of tools that enable efficient access and management of different types of information.

Although it still difficult to present an unbiased historical perspective, some authors have offered a personal view of this field, for example Trifonov (2000) and Ouzounis and Valencia (2003).

1.2 Molecular biology and genetics

This section describes some molecular cell biology basic notions useful as a background to the next chapters and provides a general biological motivation for the present work on sequence analysis. A brief introduction to biological sequences – DNA, RNA and proteins – is included, along with major cell genetic-related mechanisms explanation, such as protein synthesis and information flow, illustrated by the central dogma.

The focus is given to the eukaryote cell (characteristic of Animals, Plants, Protists and Fungi – or Eukarya), whose distinctive characteristic is a membrane enclosing the nucleus and organelles. Some of its features are shared by the prokaryote cell (Bacteria), although some processes are absent, for example RNA splicing.

1.2.1 Biological sequences

The fact of biological sequences – DNA, RNA and Proteins – being involved in the most important cell processes[4] has led to a growing interest in their analysis, with different approaches arising from various scientific fields.

These molecules have a fundamental role, defining almost all cell's activities. To present just an example, in multicellular organisms all cells have the same genetic material nonetheless they can express completely different proteins and perform distinct tasks, exhibiting an extremely diverse behavior. The key to understand these phenomena lies in comprehending the way these sequences interact with each other and with the environment that surrounds them.

[3] NCBI primer - http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html

[4] We abusively use sequences and the molecules they represent interchangeably, although the difference should be clear.

DNA and RNA

Deoxyribonucleic acid (DNA) is the basic information macromolecule of cells. It is constituted by two chains of *nucleotides*, which are composed by deoxyribose, a pentose or five-carbon sugar molecule, linked to a phosphate group and to a nitrogen organic base of one of four types: adenine (A), guanine (G), cytosine (C) and thymine (T).

Ribonucleic acid (RNA) is a single nucleotide strand exhibiting a similar composition, but with a different constituent sugar – the ribose – and uracil (U) instead of the thymine base.

In Table 1.1 DNA and RNA symbols are summarized, along with the corresponding abbreviations in the extended IUPAC (International Union Of Pure And Applied Chemistry) alphabet, commonly used in main applications.

Nucleic Acid	**Symbol**	**Meaning**
Adenosine	A	
Cytosine	C	
Guanine	G	
Thymine	T	
Uracil	U	
puRine	R	A or G
pYrimidine	Y	C or T
Weak interaction	W	A or T
Strong interaction	S	C or G
Keto	K	G or T
aMino	M	A or C
not-T	V	A or C or G
not-G	H	A or C or T
not-C	D	A or G or T
not-A	B	C or G or T
aNy	N (X)	G or A or T or C

Table 1.1: IUPAC extended DNA/RNA alphabet. Bases name, symbols and meaning. The X symbol usually refers to an *unknown* and N to an *unspecified* nucleotide.

Nucleotides in each DNA chain are connected by a chemical bond between the sugar of one nucleotide and the phosphate group of the adjacent one. When two DNA strands establish hydrogen bonds between their bases, with standard Watson-Crick pairing A—T and C—G, the classic double-helix is formed, a stable 3-dimensional structure – see Fig. 1.2.

A linear double-stranded DNA molecule and associated proteins constitute a *chromosome* and the total DNA in the chromosomes of an organism is referred to as its *genome*. The human genome has about $3 \cdot 10^9$ base pairs distributed along 46 chromosomes. The majority of human DNA has unknown function, the so-called "junk" DNA. The other part is constituted by *genes*, which are the units of hereditary information and specify the synthesis of a single polypeptide chain. Human genome is though to have around 30 000–40 000 genes. Genes are organized in exons and introns (see following sections for more details).

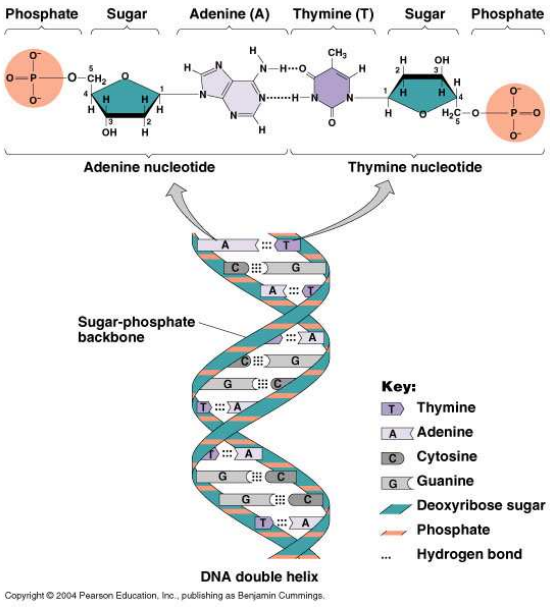


Figure 1.2: DNA structure and composition, formed by two complementary antiparallel chains of nucleotides. In Tortora et al. (2004) – used with permission © Pearson Education, Inc.

DNA sequences are often represented using a 4-symbol alphabet that transcribes the coding or sense strand, from the 5' to the 3' end (referring to the free carbon in the terminal sugar).

Different types of RNA are involved in distinct cell processes, namely messenger (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA), described below.

Protein structure and function

Proteins are macromolecules made of tri-dimensional polypeptide[5] chains of *aminoacids* (aa). All 20 aminoacids share a general structure – they are constituted by one carbon atom bonded to four different chemical groups: amino (NH2), carboxyl (COOH), hydrogen (H) and a side chain (R) group that defines their name and distinct biochemical properties. Table 1.2 summarizes the aminoacids names and abbreviated symbols, referred to the side chain group R.

Aminoacids in a protein chain are connected by peptide bonds, formed by a chemical reaction between the amino group of one amino acid and the carboxyl group of another.

Proteins carry out most of the cell biological activities and are encoded by genes. Their role is of vital importance in all processes, from all regulatory functions, acting as biological switches, to signaling and intermembranar transport between the interior and exterior of the cell. It is noteworthy that all the biochemical reactions are catalyzed by *enzymes*, that constitute a large protein category.

[5] A peptide is a chain with less than 50 aa.

Aminoacid name	Symbol	Abbreviation
Alanine	A	Ala
Arginine	R	Arg
Asparagine	N	Asn
Aspartic Acid	D	Asp
Cysteine	C	Cys
Glutamine	Q	Gln
Glutamic acid	E	Glu
Glycine	G	Gly
Histidine*	H	His
Isoleucine*	I	Ile
Leucine*	L	Leu
Lysine*	K	Lys
Methionine*	M	Met
Phenylalanine*	F	Phe
Proline	P	Pro
Serine	S	Ser
Threonine*	T	Thr
Tryptophan*	W	Trp
Tyrosine	Y	Tyr
Valine*	V	Val

Table 1.2: Aminoacid alphabet – names and symbols. The star symbol * represents essential aminoacids, which cannot be synthesized by human cells.

Protein structure is of fundamental importance to define its function, since all processes occur in a 3-dimensional space and require a specific configuration for attaining a precise aim. Therefore the development of methods to predict their spatial arrangement is of major importance and has become recently a strong research topic in bioinformatics, yet with not a definitive answer, with accuracies far from 100%.

Proteins have four hierarchical levels of structure. The protein *primary structure* is the linear arrangements or sequence of its aminoacids. The *secondary structure* corresponds to local organization. When hydrogen bonds are created between residues, two structures become apparent: the α-helix, a spiral conformation, and the β-sheet, a planar structure. The *tertiary structure* is the full tri-dimensional folded arrangement or overall conformation of the polypeptide chain. Finally, the *quaternary structure* appears when more than one protein polypeptide chains are held together, creating complex interconnections. Figure 1.3 presents an example of the four hierarchical levels described, showing the hemoglobin structure. This example will resume later in this work.

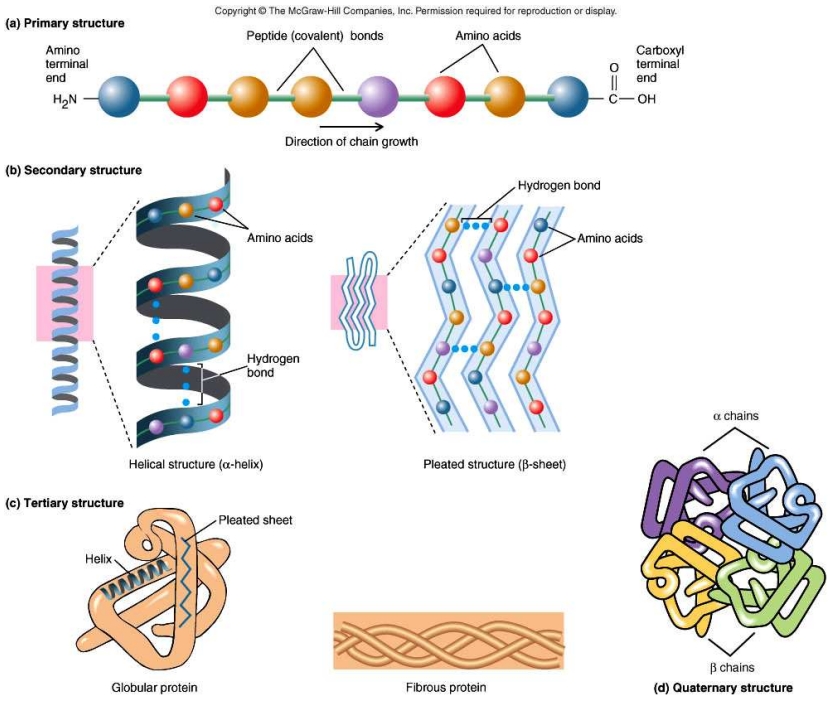


Figure 1.3: Protein hierarchical structure of hemoglobin . a) Primary structure – aminoacid sequence; b) Secondary structure – local folding; c) Tertiary structure – long-range folding; d) Quaternary structure – multi-chain organization. In Nester et al. (2004) – used with permission © McGraw-Hill.

1.2.2 Synthesis of macromolecules and the central dogma

The main molecules involved in cell mechanisms have been described previously; this section is devoted to the explanation of how the information is passed from genes to genes and from genes to proteins. Figure 1.4 depicts some of the mechanisms that occur in an eukaryotic cell and that will be briefly reviewed. The

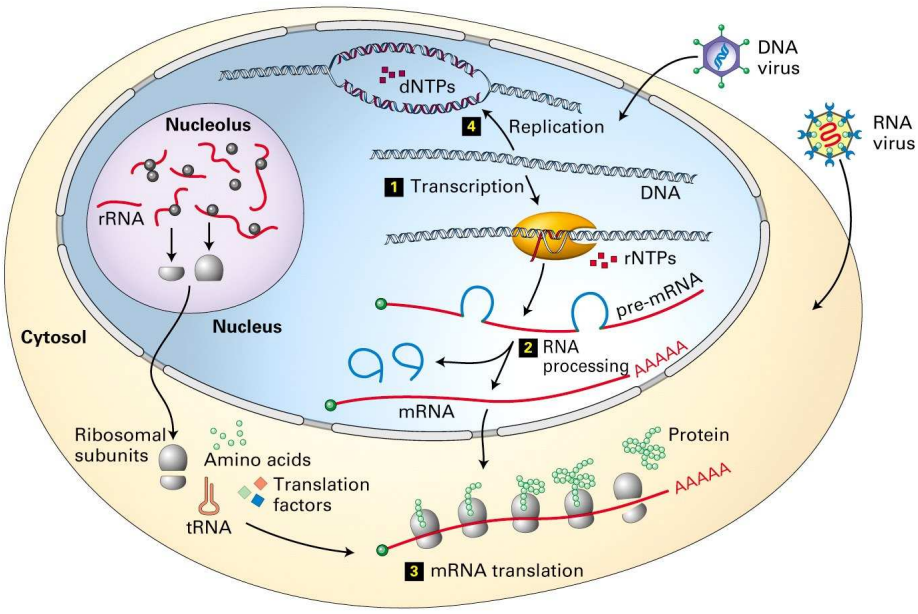


Figure 1.4: Macromolecule synthesis and other eukaryote cell mechanisms. Protein synthesis (steps 1–3) and DNA replication (step 4) described in the text. In Lodish et al. (2004) – used with permission © W.H. Freeman and Company.

central dogma states that information flow is from DNA → RNA → proteins, according to the processes described below.

DNA replication

When cells divide, all genetic information stored in genes is duplicated. This is accomplished by the synthesis of a new DNA molecule using the older one as a template, from the 5' to the 3' end. In this process both strands are duplicated and eventually will belong to two new different DNA chains. The *replication* – see Fig. 1.4 – is done with high accuracy, with very precise error correction mechanisms due to the repair effect of the enzyme DNA polymerase, at a rate that varies from 500 to one million base pairs *per* minute in bacteria.

Protein synthesis

Protein synthesis is a fundamental process in which the information encoded in DNA is expressed and effectively pass and influence the cell structure and metabolism. It involves several steps, also mediated by enzymes, and where different types of RNA perform an important role (see Fig. 1.4).

1st \ 2nd	U	C	A	G	3rd
U	Phe F	Ser S	Tyr Y	Cys C	U
					C
	Leu L		*stop*	*stop*	A
			stop	trp W	G
C	Leu L	Pro P	His H	Arg R	U
					C
			Gln G		A
					G
A	Ile I	Thr T	Asn N	Ser S	U
					C
			Lys K	Arg R	A
	Met M				G
G	Val V	Ala A	Asp D	Gly G	U
					C
			Glu E		A
					G

Table 1.3: The genetic code, written by convention in the form in which the codons appear in mRNAs. Equivalence of DNA codons and aminoacids during translation. Each codon corresponds to 3 bases in positions (1st, 2nd, 3rd). The first symbol in the codon corresponds to the left column. Example: the codon AGU corresponds to Ser (S). The codon AUG corresponds to Met (M) and also to the *start* or initiator codon, i.e., it signals the beginning of transcription – this aminoacid might be removed afterwards

The first step is the *transcription* where a precursor-mRNA (pre-mRNA) molecule is synthesized using DNA as a template and forming a complementary strand: A–U, T–A, C–G, G–C. In eucaryote DNA this molecule is further processed by *splicing*, where the *introns* (intervening sequences) are excised and the *exons* (expressed sequences) are maintained and linked, forming a mRNA chain containing the filtered information for protein synthesis. Only this completed and mature mRNA is selectively transported from the nucleus to the cytoplasm. Alternative splicing increases coding potential of genomes since it produces, from the same DNA sequences, different mRNA chains.

This mRNA further attaches to the ribosome, which consists of ribosomal RNA (rRNA) and proteins, and the *translation* process begins. In this step the information contained in the mRNA is decoded into proteins, i.e., aminoacids are added one at a time, from the amino (NH2) to the carboxyl (COOH) terminus, following the mRNA template. The key step of translation is the rule code associated, in which each group of 3 nucleotides – so-called *codon* – specifies one aminoacid. Table 1.3 represents the *genetic code*, i.e., the correspondence between each possible codon – defined by 3 symbols (1st, 2nd, 3rd) – and each aminoacid.

Since there are $4^3 = 64$ codons and only 20 aminoacids there is some re-

dundancy in the translation and often only the first two symbols of a codon are needed to univocally specify an aminoacid – the code is said to be *degenerate*. Moreover, there are special codons that define the beginning and the end of translation: the *start* codon (AUG) and the *stop* codons (UAA, UAG and UGA). Depending on the starting point – 1st, 2nd or 3rd symbol – there are three different *reading frames* associated with one DNA chain, that will built different proteins.

The translation process, as previously seen, begins with the recognition of the start codon by the ribosome and its subsequent attachment to the mRNA. The transfer RNA (tRNA) is the key molecule that will perform the connection between bases and aminoacids. The tRNA is a type of RNA that has a 3-base sequence, or *anti-codon*, that can pair with its complementary code in the mRNA and also can bind and carry a specific aminoacid, following the rules of the genetic code on Tab. 1.3. The translation process is shown in the Fig. 1.5 and recalled below.

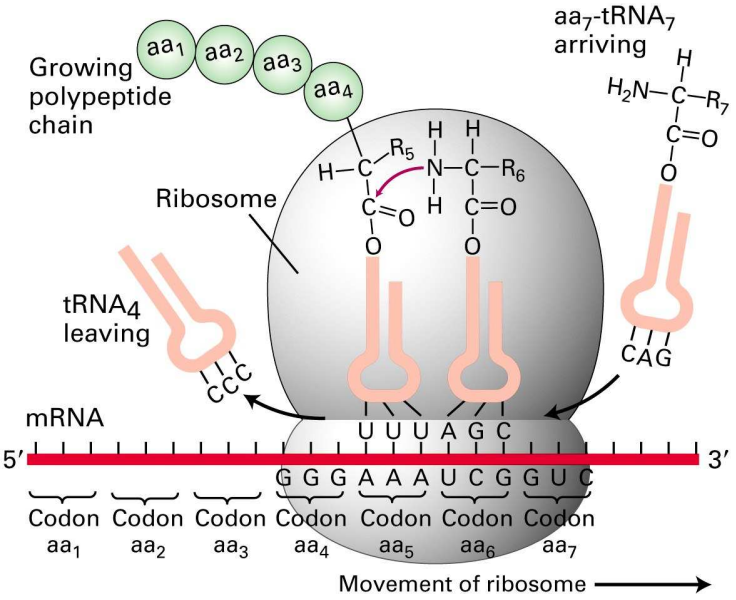


Figure 1.5: Protein synthesis and RNA roles in translation. This figure shows the translation step of protein synthesis, in which the ribosome (rRNA + proteins) binds to the start codon of mRNA, which in turn binds to the complementary triplet of tRNA, that sequentially carries a new aminoacid to the polypeptide chain. In Lodish et al. (2004) – used with permission © W.H. Freeman and Company.

The tRNA carrying a specific aminoacid binds to the complementary mRNA codon. The next mRNA codon in the sequence defines another complementary tRNA that brings the corresponding aminoacid, that will form a peptide bond with the first one. The elongation processes continues, with new aminoacids forming bonds with the growing polypeptide chain, the tRNA is released and the ribosome shifts to the next codon available. The process ends when the ribosome reaches a stop codon, releasing the mRNA and the new protein just

synthesized.

After this step the protein folds onto a 3-dimensional conformation and might also undergo some post-translational modifications that alter its structure and function.

1.3 Sequence analysis and comparison

Sequence analysis, the main subject of this thesis, is in the core of almost all bioinformatics applications (Durbin et al., 1998). Even with the recent explorations of higher-level integrative data, such as microarrays and genetic regulatory networks, sequence analysis and comparison is still a vital subject since almost all tasks depend on algorithms that process and investigate strings, from searching for similar sequences in databases to classification problems.

Although the present work is centered in vector mappings of biological sequences – *alignment-free* methods – a brief introduction to alignment algorithms is warranted to fully recognize the different approaches to sequence analysis and how they are related to each other.

Biological sequences are usually represented as strings whose symbols belong to the alphabets described in the last section – bases for DNA (Tab. 1.1) and aminoacids for proteins (Tab. 1.2). An introduction to words in sequences is given in section 2.2.1 on page 36 (Vinga and Almeida, 2003) and Gusfield (1997) describes algorithms on strings typically used in bioinformatics.

1.3.1 Alignment methods for sequence comparison

Since the beginning of bioinformatics alignment methods have been used extensively, based on the premise that if two sequences share some substrings, accepting some degree of mutations or 'errors', they might also have a common ancestor and similar function. This simple approach is in the origin of several algorithms whose objective is to 'optimally' align two sequences, to uncover their presumed common root.

Motivation

As DNA is duplicated and passed to the next generations, some permanent changes can occur that might lead to structure modifications of the resultant encoded protein and subsequent alteration of its function, a process called *mutation*. These mutations can occur for several reasons, from errors in the duplication to the exposure to environmental factors, such as virus, chemical agents or radiation. There are several types of mutations; the simplest one is a *single point mutation* when one single base is substituted by another. One example of a point mutation with serious impact in human health is the substitution of a nucleotide in the β-globin gene, which codifies for hemoglobin, further explained in section 2.4.3 on page 49. Other possible alterations are *insertions*, where one or more bases are added in the original DNA and *deletions*, in which a piece of DNA is excised. As these mutations get passed along descendants, the original sequences will diverge, at different rates. Sequences that have a

common ancestor are called *homologous*. These variations are the key to evolutionary processes and natural selection, since they amplify the possible range of individuals and phenotypes and the more favorable ones tend to accumulate, leading to the survival of the best fit.

A *pairwise alignment* is simply an arrangement of two sequences, one on top of the other, highlighting their common symbols and sub-strings, represented as vertical *matches* "|". The alignment can also have mismatches, i.e., a symbol in one of the sequences is different from the other, which corresponds to a single point mutation; if the residues are 'similar' in some sense, the symbol ":" is used. Gapped alignments also permit *in*sertions or *del*etions – or *indels* – while n consecutive indels constitute a *gap* of length n, represented as "–". In theory, alignments allow to trace back the mutations, given putative homologous sequences.

The fundamental idea of alignment is that sequences that share the same substrings might have the same function or be related by homology. To some extent, this naive procedure was applied for deciphering the hieroglyphs, from the Rosetta stone information. This stone had the same text written in three different scripts: in hieroglyphic (ancient Egyptian) demotic (cursive and more recent Egyptian) and Greek. Assuming that the information (function) was the same along these three sequences, i.e., once known that the message was the same across the texts, it was possible for Jean-François Champollion to deduce all the code encrypted in the hieroglyphs and to establish the foundation of modern Egyptology.

Algorithms and scoring schemes

Alignment algorithms are anchored in specific scoring schemes. These include *substitution or scoring matrices*[6], which assign a value for aligning each two symbols in the alphabet and reflects how conservative is the substitution between them, and *gap score penalties* $\gamma(n)$, a negative quantity that penalizes for n consecutive indels. The total score of an alignment will be the sum of terms for each aligned pair of residues – extracted from the scoring matrix – plus terms for each gap.

The global alignment computational problem can now be stated as follows: what is the optimal alignment of two sequences, i.e., the one that has the highest score, for a particular scoring scheme. A naive procedure would be to perform all possible alignments between the two sequences and choose the one that maximizes the score. Of course this would be computationally unfeasible, even for short sequences.

The solution was established by Needleman and Wunsch (1970) and computationally improved afterwards. The algorithm uses *dynamic programming*, which is a technique based on breaking up the main problem in simpler subproblems, that in turn can be further divided. The optimization of the smaller problems, *recursively*, will eventually lead to the optimal solution of the main problem.

In this case, finding the best alignment between two sequences $x_1 \cdots x_n$

[6] PAM and BLOSUM matrices are used for protein alignment – see page 86.

and $y_1 \cdots y_m$, with total score $F(n,m)$ uses the solution of smaller alignment problems, namely the best alignment between prefixes of $x_1 \cdots x_i$ and $y_1 \cdots y_j$, with score $F(i,j)$, $i = 1,\ldots,n$ and $j = 1,\ldots,m$.

As an example, Fig. 1.6 shows the alignment of two sequences $\mathbf{S_1}$ = ATCGCCAAT and $\mathbf{S_2}$ = ATGCCGCCT. The scoring matrix chosen as-

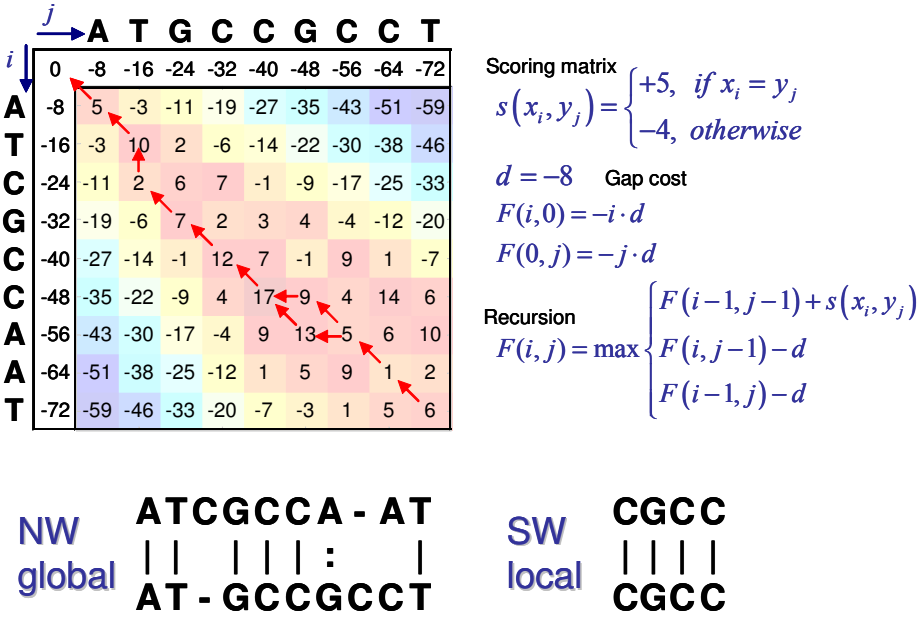


Figure 1.6: Pairwise alignment example

signs the value $s(x_i,y_j) = +5$ to a match $x_i = y_j$ and $s(x_i,y_j) = -4$ to a mismatch $x_i \neq y_j$. The scoring function for gaps used was a *linear gap penalty* $\gamma(n) = d \cdot n = -8n$, i.e., a cost of 8 units per gap character.[7] The recursion procedure is also explicit in the figure: the initialization step calculates all the values in the first row $F(0,j)$ and the first column $F(i,0)$ of the matrix, as if the alignment would be only constituted by a gap. In each of the following steps three quantities are calculated that are equivalent to three different options for the alignment: align the two following symbols has a cost of $F(i-1,j-1) + s(x_i,y_j)$, open a gap in the first sequence or open a gap in the second, with scores $F(i,j-1) + d$ and $F(i-1,j) + d$ respectively.

The key is to solve the sub-problem in this step, by choosing the option that maximizes the partial score $F(i,j)$, i.e., the optimal solution for those prefixes, also keeping track of that solution. After filling all the matrix F following this procedure, one can *traceback* the path, from $F(n,m)$ to the beginning of the sequences $F(0,0)$, following the indexes that originated the maxima – represented with arrows in the figure. Finally, the optimal global alignment is extracted – shown below the matrix. It is guaranteed that with this stepwise procedure the optimal alignment, i.e., the one with the highest score, will be

[7] Alternatively, *affine gap penalties* have the form $\gamma(n) = d \cdot (n-1) + a$, introducing a gap-open cost a and a gap-extension penalty d.

obtained, in this case $F(9,9) = 6$.

The *local* alignment of the two sequences, also presented in Fig. 1.6, identifies the best sub-string alignment, ignoring low-scoring parts in the sequences. It was obtained with the algorithm developed by Smith and Waterman (1981), which assigns different boundary conditions and slightly modifies the recursion to avoid negative numbers in the matrix. This algorithm was used to compare alignment-based dissimilarities with word composition distances in Vinga et al. (2004), transcribed in Chapter 4.

Statistical significance

The correct evaluation of the statistical significance of alignment scores has been the subject of several studies. The goal is to determine if the value obtained with an optimal alignment does in fact translate onto biological similarity or might have occurred just by chance. One possible approach uses Bayesian statistics, and calculates *a posteriori* probabilities of the sequences being related. The other approach uses asymptotic theory and extreme value distribution to model the cumulative probability function of obtaining a particular score. For more complex models and algorithms using heuristics this is still an open problem.

Applications and software

Among the applications of these methods are the ubiquitous search of sequences in databases, that allow for heuristic procedures to speed the process. Other applications include the construction of phylogenetic trees from dissimilarities obtained from alignment scores, modelled with evolutionary assumptions. Specially known procedures are BLAST – Basic Local Alignment Search Tool (Altschul et al., 1990, 1997) and FASTA (Pearson and Lipman, 1988; Pearson, 1990). For multiple alignments, CLUSTAL (Thompson et al., 1994) is the most commonly used software – for an example of the output, see Fig. 2.6 on page 54. Nowadays, several resources provide downloadable versions of these toolboxes and/or make them available online.[8]

1.3.2 Vector maps for sequence comparison

This work is centered on alternative representations of biological sequences that do not rely on alignment methods but in vector maps. A *vector map* or *mapping* is a vector-valued function, i.e., a function that assumes values on the vector space $\mathbb{R}^n$. Given any sequence $\mathbf{S} = s_1 s_2 \cdots s_i \cdots$, from an alphabet $s_i \in \mathcal{A}$, we are interested in functions $f(\mathbf{S}) \in \mathbb{R}^n$:

$$f : \mathbf{S} \rightarrow \mathbb{R}^n \tag{1.1}$$

In this thesis two types of functions will be studied. Both transform symbols and sequences into vectors in $\mathbb{R}^n$. These two methods, along with several others not exploited in this work, were reviewed by Vinga and Almeida (2003) and are presented in Chapter 2.

[8] For example http://www.ebi.ac.uk/Tools/

The first type is based on mapping a full sequence onto its substring frequency vectors, thus obtaining its word composition. By comparing these vectors it is possible to define several dissimilarity measures suitable for specific applications, such as proteins classification (Vinga et al., 2004).

The second type of function is based on a different approach in which, instead of mapping all the sequence, a function is applied to each of its symbols iteratively. The algorithm beneath – chaos game representation (CGR) – has its roots in fractal theory. The full description of this technique is presented in section 1.4.2 (p. 17) and its generalization – Universal Sequence Maps (USM) – is fully exploited in Chapter 3 (Almeida and Vinga, 2002). One application of this representation is the estimation of DNA entropy, presented in Chapter 5, where the work by Vinga and Almeida (2004) is transcribed.

1.4 Iterative Function Systems

This section presents an introduction to iterative function systems (IFS) and fractal geometry[9], from which chaos game representation (CGR) and Universal Sequence Maps (USM) are derived. CGR/USM algorithms are explored in this work as alternative methods to represent and analyze biological sequences. In particular, the new measure of DNA uncertainty (or entropy) presented in Chapter 5 is anchored in this mapping (Vinga and Almeida, 2004).

The most part of this introduction is fully extended in (Barnsley, 1998) and (Edgar, 1990).

1.4.1 Definitions

A metric space (X, d) is a set X together with a function $d : X \times X \to \mathbb{R}_0^+$ satisfying:

$$\begin{aligned} d(x,y) = 0 \Leftrightarrow x = y \\ d(x,y) = d(y,x) \\ d(x,y) \leq d(x,z) + d(z,y) \end{aligned} \tag{1.2}$$

Such a function d is called a *metric* and measures the distance between pairs of points x and y in X.

For example the function $d^{EU}(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$ is a metric in the set $X = \mathbb{R}^2$. This means that $(\mathbb{R}^2, d^{EU})$ is a metric space.

A transformation $f : X \to X$ on a metric space (X, d) is called a *similarity* if and only if there is a positive number r such that

$$d\left(f(x), f(y)\right) = r \cdot d(x,y), \qquad \forall x, y \in X \tag{1.3}$$

The number r is called the *ratio* of f.

[9] Fractals are geometric objects that can be generated by different methods, from recurrence relations to stochastic processes, but in this introduction they will be exclusively described as attractors of IFS.

When $0 \leq r < 1$ the function f is also called contractive or a *contraction mapping*, since in this case the distances decrease by a factor r that is less than 1.

An *iterated function system* (IFS) consists of a complete metric space[10] (X, d) together with a finite set of contraction mappings $f_i : X \to X$, with respective ratios r_i, for $i = 1, 2, \ldots, n$.

The *dimension* associated with $\{r_i\}_{i=1,\ldots,n}$ is the positive number s such that $r_1^s + r_2^s + \cdots + r_n^s = 1$. It can be shown that s always exists.

A nonempty compact set $A \subseteq X$ is an *invariant set* or *attractor* of an IFS $(f_1, \ldots, f_n)$ if and only if $A = f_1(A) \cup \cdots \cup f_n(A) = \bigcup_{i=1}^n f_i(A)$. This means that A is mapped onto itself by the IFS. It can be proven that there is a unique nonempty compact invariant set A for one particular IFS (Edgar, 1990, chap. 4). Another important corollary is that, given any nonempty compact set B_0 in X and $B_{k+1} = \bigcup_{i=1}^n f_i(B_k)$ for $k \geq 0$, the sequence $\{B_k\}$ converges to the invariant set A of the IFS. This fundamental result will be used for the construction of attractors of specific IFS. If we take *any* B_0 and iteratively apply all functions $\{f_i\}_{i=1,\ldots,n}$, keeping the union of all the sets obtained, we will obtain a series of sets that will converge to the IFS attractor A, i.e., $\lim_{k\to\infty} B_k = A$. The set thus obtained is a *fractal*.

As an example to illustrate the previous definitions, one derived IFS attractor, the *Sierpinski*[11] *triangle or gasket*, is shown in Fig. 1.7. There are several

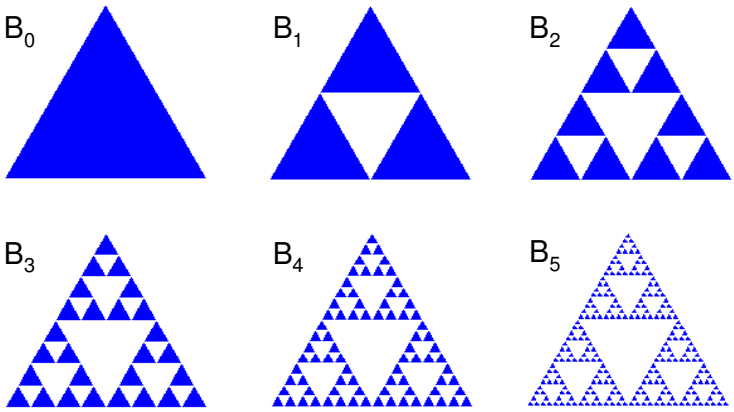


Figure 1.7: The Sierpinski triangle or gasket. This fractal is an invariant set or attractor of the IFS defined in Eq. 1.4

ways to construct this set. A deterministic construction algorithm starts with an equilateral triangle (and its boundary) in the plane, with side length 1 — B_0 in the figure. In each step, one triangle with half the side-length of the original one is removed from each of the remaining triangles, leaving the rest of the set and its boundary. For example the set B_1 corresponds to set B_0 where the middle triangle of side length 1/2 was removed, B_2 takes the 3 remaining triangles in B_1 and removes the middle triangles with side length 1/4 and so forth, always

[10] A metric space X is called complete if and only if every Cauchy sequence in X converges (in X).

[11] After Waclaw Sierpinski (Warsaw, 1882-1969).

obtaining a subset of the previous B_i, $i \in \mathbb{N}$. The Sierpinski triangle B is the limit of this decreasing sequence of sets, i.e., $B = \cap_{n\in\mathbb{N}} B_n$.

It can be proven that the Sierpinski triangle is the attractor of the IFS defined below (Eq. 1.4) consisting of 3 contraction mappings (f_1, f_2, f_3) in the plane. Each of the functions $f_i : \mathbb{R}^2 \to \mathbb{R}^2$, $i = 1, 2, 3$, have ratio or contractivity factor 1/2 and 'contraction' centers in each of the vertices of the triangle, whose coordinates are $(0, 0)$, $(1, 0)$ and $(1/2, \sqrt{3}/2)$:

$$\begin{cases} f_1(x,y) &= \frac{1}{2}(x,y) \\ f_2(x,y) &= \frac{1}{2}(x,y) + \frac{1}{2}(1,0) \\ f_3(x,y) &= \frac{1}{2}(x,y) + \frac{1}{2}\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) \end{cases} \tag{1.4}$$

This result also gives other alternatives for the construction of this set. By picking any point in B_0 and by iteratively applying the three functions (f_1, f_2, f_3) the exact same attractor will eventually be obtained. In fact, in several examples in literature the Sierpinski triangle is constructed through a process named *chaos game*, in which, starting from any point in B_0, we apply one of f_i depending on the output of a random experience, like the tossing of a 3-side (virtually) dice. This iterative process will also converge to the attractor. Intuitively, this procedure corresponds to starting in a random point in the triangle and going, in each step, half the distance towards the vertex representing the random output. This method of obtaining the attractor in which a probability p_i, with $\sum_{i=1}^{n} p_i = 1$, is assigned to each of the functions f_i is called *random iteration algorithm*.

There are some interesting remarks concerning the notions of measure and dimension associated with this type of objects. In each iteration n the set B_n consists of 3^n triangles with side 2^{-n}. So the total area of the Sierpinski set is $3^n \cdot 2^{-2n}\sqrt{3}/4$ which converges to 0 as $n \to \infty$. The total length is $3^n \cdot 2^{-n} \cdot 3$ that goes to ∞ when $n \to \infty$. This example shows that neither length or area are useful in the description of B measure. On the other hand, the above defined dimension s of the Sierpinski triangle can be obtained from the ratios of the contraction mappings, as the solution of $\sum_{i=1}^{3} r_i^s = 1$. Since the ratio list is $(1/2, 1/2, 1/2)$ the dimension is $3 \cdot (1/2)^s = 1$, thus $s = \log 3/\log 2 \approx 1.585$. This number is between 1 and 2, what is in agreement with the previous perception of its dimension being between 'length' and 'area'.

Along with the above presented features (its non-integer dimension and the iterative procedure taken to obtain this set), the Sierpinski triangle also illustrates other properties of fractals such as *self-similarity*, which means that it appears identical at different scales. When magnifying this set, for example the upper triangle in B_1, we get an exact replica of the whole Sierpinski triangle.

1.4.2 Chaos game representation (CGR)

Chaos game representation (CGR) was first presented in 1990 (Jeffrey, 1990) as a method for representing DNA sequences on vectorial spaces. It is derived from IFS theory, briefly described in the last section. There are several applications of this method in bioinformatics, such as the investigation of patterns in DNA, the

extraction of Markov models transition tables and the calculation of entropies. The algorithm itself is closely related to binary representations of sequences and suffix trees. The following sections briefly describe the main properties and results of this method. The CGR generalization for higher-order alphabets is fully presented in Chapter 3 (Almeida and Vinga, 2002).

CGR definition

The iterative algorithm is constructed on a square in $\mathbb{R}^2$ where each of its vertex is assigned to a DNA symbol or base ($\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}$) – see section 1.2. For a given DNA sequence $\mathbf{S} = s_1 s_2 \cdots s_N$ with length N, $s_i \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}, i = 1, \ldots, N$, CGR maps each symbol s_i onto a point $x_i \in \mathbb{R}^2$ following an iterative procedure defined in Eq. 1.5. The original proposal assigned the first point x_0 to the center of the square $[0, 1]^2$.

$$\begin{cases} x_0 = (0.5, 0.5) \\ x_i = x_{i-1} + \frac{1}{2}(y_i - x_{i-1}), i = 1, \ldots, N \end{cases} \quad \text{where} \quad y_i = \begin{cases} (0,0) & \text{if } s_i = \mathsf{A} \\ (0,1) & \text{if } s_i = \mathsf{C} \\ (1,0) & \text{if } s_i = \mathsf{G} \\ (1,1) & \text{if } s_i = \mathsf{T} \end{cases} \tag{1.5}$$

An alternative formula for one particular x_i representing the i$^{\text{th}}$ symbol as a function of all previous positions is given by:

$$x_i = 2^{-i} x_0 + \sum_{k=1}^{i} 2^{-k} y_{i-k+1} \quad , \quad i = 1, \ldots, N \tag{1.6}$$

In Figure 1.8 an example of a CGR map is shown, using the human hemoglobin gene sequence. This example will be further extended in section 2.4.3.

This mapping can be interpreted as an IFS on the square $[0, 1]^2$ in which a random iteration algorithm is being used. There are 4 contraction mappings (f_1, f_2, f_3, f_4), one for each DNA base. Each of the functions f_i have ratio 1/2 and 'contraction' centers in each vertex, one per symbol, and is specified by symbol s_i in the original DNA sequence. This procedure is very similar to the random construction of the Sierpinski triangle described above.

CGR properties

One important property of chaos game representation is the 'closeness' of points in the space $[0, 1]^2$ when the symbols they represent are the same. In fact, the CGR maps same substrings close to each other, which means that, wherever the context is, the same suffix will be always mapped in the same region of CGR.

Let us suppose we have two sequences $\mathbf{S} = s_1^S s_2^S \cdots s_N^S$ and $\mathbf{T} = s_1^T s_2^T \cdots s_M^T$ that in a given position share the same suffix of length L, $s_{i+1}^S \cdots s_{i+L}^S = s_{j+1}^T \cdots s_{j+L}^T$ for some $i \in \{0, \ldots, N - L\}$ and $j \in \{0, \ldots, M - L\}$. The CGR coordinates of these points are calculated with CGR iteration previously defined (Eq. 1.5) with $y_{i+k}^S = y_{j+k}^T$ through all the shared string $k = 1, \ldots, L$. In this step we are interested in what happens to the distances between these

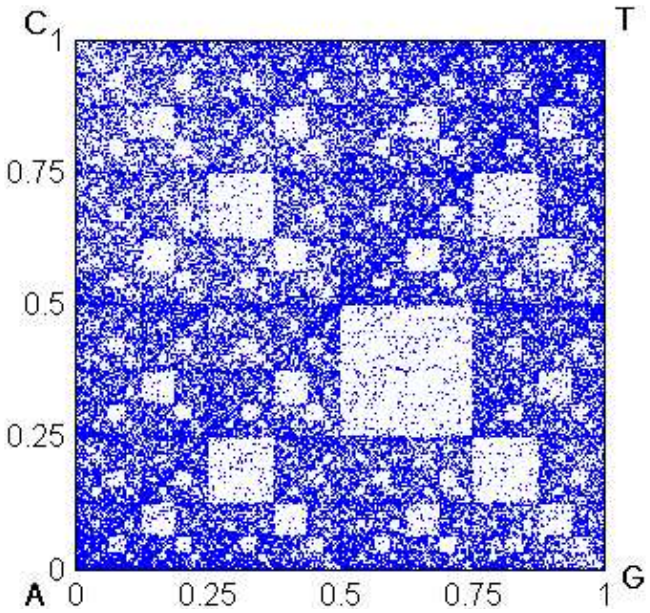


Figure 1.8: CGR of human beta globin region on chromosome 11 (HUMHBB) - 73308 bases.

coordinates, i.e., to the quantity $x^S_{i+k} - x^T_{j+k}$, for $k = 1, \ldots, L$. Starting from the coordinates x^S_i and x^T_j we have:

$$\begin{aligned} x^S_{i+1} - x^T_{j+1} &= 2^{-1}(x^S_i - x^T_j) \\ x^S_{i+2} - x^T_{j+2} &= 2^{-1}(x^S_{i+1} - x^T_{j+1}) = 2^{-2}(x^S_i - x^T_j) \\ &\ldots \\ x^S_{i+L} - x^T_{j+L} &= 2^{-L}(x^S_i - x^T_j) \\ \nabla^L x &= 2^{-L} \cdot \nabla^0 x \end{aligned} \tag{1.7}$$

The last formula shows that the difference between coordinates is decreased by a factor of 2 in each common symbol. One consequence of this result is that the CGR map can be divided and labelled according to the corresponding substring, i.e., each substring is mapped onto a sub-square. An example of this property is shown in Fig. 5.1 on page 105.

CGR and binary numbers

The CGR algorithm can also be interpreted as binary expansions of numbers and operations performed in base 2. Normally the base used is 10 and is omitted in most of the representations. Base 2 works in a similar way, where each positive number x has a unique representation $x = (a_M a_{M-1} \cdots a_1 a_0 . a_{-1} \cdots)_2 = \sum_{j=-\infty}^{M} a_j 2^j$, with binary digits $a_j \in \{0, 1\}$, also called *bits*. For example the number 5.25 is written, in base 2, as $5.25 = (101.01)_2 = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2}$.

Some mathematical operations have straightforward results, for example, dividing by two is simply $x/2 = \sum_{j=-\infty}^{M} a_j 2^{j-1}$, which corresponds to the shifting of all digits one position to the left, i.e., the digit a_j originally associated with position j, power 2^j, after division will be the digit of position $j-1$, power 2^{j-1}. In the former example $5.25/2 = (10.101)_2 = 1 \cdot 2^1 + 0 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} = 2.625$.

This particular operation relates to CGR in the sense that, in each iteration, the contraction function has ratio 1/2. For example, when representing the sequence $\mathbf{S} = \mathsf{CTAG}$ starting from $x_0 = (0.5, 0.5) = (0.1, 0.1)_2$, the next points $\{x_i\}_{i=1,\ldots,4}$ will be obtained from the contraction mappings f_i as:

$$\begin{aligned}
x_1 &= \frac{1}{2}(0.5, 0.5) + \frac{1}{2}(\mathbf{0}, \mathbf{1}) \\
&= \frac{1}{2}(0.1, 0.1)_2 + \frac{1}{2}(0, 1)_2 \\
&= (0.01, 0.01)_2 + (0.0, 0.1)_2 \\
&= (0.\mathbf{0}1, 0.\mathbf{1}1)_2 \\
x_2 &= \frac{1}{2}(0.01, 0.11)_2 + \frac{1}{2}(\mathbf{1}, \mathbf{1})_2 \\
&= (0.001, 0.011)_2 + (0.1, 0.1)_2 \\
&= (0.\mathbf{1}01, 0.\mathbf{1}11)_2 \\
x_3 &= (0.\mathbf{0}101, 0.\mathbf{0}111)_2 \\
x_4 &= (0.\mathbf{1}0101, 0.\mathbf{0}0111)_2
\end{aligned} \tag{1.8}$$

The process is now clear: in each step the previous number is shifted one bit and a new number is added in the form $(0.a, 0.b)$, where (a, b) corresponds to each of the vertex/symbols.

In this format is also easy to prove the bijective property of CGR – knowing a given point x_k it is possible to recover all the sequence up to that position.

CGR and Markov chains

Due to CGR mapping properties there is a relationship between this representation and Markov chains (MC), a special case of a stochastic model. This result was presented in (Almeida et al., 2001), showing that CGR correctly accommodates MC models.

A *stochastic process* is a collection of random variables $\{X(\alpha), \alpha \in T\}$, indexed by the parameter α taking values in the parameter set T (Kulkarni, 1995). The random variables $X(\alpha)$ take values in a state space S. In several bioinformatics applications a sequence is modelled as a stochastic process, in which the state space S is an alphabet $\mathcal{A}$ – bases for DNA and aminoacids for protein as in Tab. 1.1 and 1.2 – and the parameter α is the position i of each symbol in the sequence $\{X_i, i \in \mathbb{N}\}$, with $X_i \in \mathcal{A}$.

The probability of observing a given finite sequence $\mathbf{S} = s_1 s_2 \cdots s_i \cdots s_N$ with $s_i \in \mathcal{A}$ is represented as:

$$p(X_N = s_N, \ldots, X_i = s_i, \ldots, X_2 = s_2, X_1 = s_1) = p(s_1 s_2 \cdots s_i \cdots s_N) \tag{1.9}$$

A special type of models widely used in sequence analysis is the Markov chain. A *Markov chain* is a stochastic process with a 'memoryless' or Markovian property that can be stated as follows: if the present of the system is known, its future is independent from the past. Formally, this relation can be expressed with conditional probabilities.

When applied to sequence modelling, this means that knowing a specific symbol in one position, the probabilities of the following symbol are independent from the previous ones, i.e., the present contains all the relevant information needed to predict the future.

Using the Markov property just described, the probability of observing symbol $s_{i+1} \in \mathcal{A}$ on position $i+1$, given all the observed sequence $s_1 s_2 \cdots s_i$ depends only on the last observed symbol s_i:

$$\begin{aligned} p(X_{i+1} = s_{i+1} | X_i = s_i, X_{i-1} = s_{i-1}, \ldots, X_1 = s_1) &= p(X_{i+1} = s_{i+1} | X_i = s_i) \\ \Leftrightarrow p(s_{i+1} | s_1 s_2 \cdots s_i) &= p(s_{i+1} | s_i) \end{aligned} \tag{1.10}$$

The last equation uses a notation simplification that will be used further on, otherwise noted, where the variable X is omitted. The value $p(s_{i+1}|s_i)$, also represented as $p_{s_i s_{i+1}}$, is called the *transition probability* from s_i to s_{i+1}. For *homogeneous* Markov chains, this probability does not depend on the specific position i considered, being the same all through the sequence. Therefore, it is possible to rearrange all these conditional probabilities in one *transition probability matrix* of dimension $|\mathcal{A}| \times |\mathcal{A}|$, where $|\mathcal{A}|$ is the alphabet length.

One natural extension of this idea uses longer-memory Markov models, in which, instead of considering just the last symbol s_i, a suffix of length L is taken $s_{i-L+1} \cdots s_i$.

$$p(s_{i+1} | s_1 s_2 \cdots s_i) = p(s_{i+1} | s_{i-L+1} \cdots s_i) \tag{1.11}$$

More formally, a L-order Markov chain is characterized by conditional distributions or transition probability tables, which define the probability of one state given the current L-tuple or suffix ending in the present position.

Using CGR it is possible to estimate all the transition probabilities by extracting the number of points in each quadrant. This is strongly related to the suffix property described above, stating that each suffix is mapped onto a specific CGR sub-quadrant of size $2^{-L} \times 2^{-L}$. For example, if the sequence is modelled as a $L = 1$ order Markov chain, by extracting all di-nucleotides counts it is possible to estimate the transition probabilities, as illustrated in Fig. 1.9.

More generally, in order to extract the transition probabilities estimates for an L-order Markov chain, the interval $[0, 1]$ should be divided in 2^{L+1} sub-intervals. The number of points in each sub-square thus created is then counted, which is the same as extracting the number of $(L + 1)$-nucleotides.

More recently CGR was further explored, using the above mentioned properties, for time series prediction by fractal prediction machines (FPM) (Tino and Dorffner, 2001), showing better performances than other models, e.g. variable length Markov models (VLMM). These results demonstrate that CGR can be used as generalization of Markov chain models.

CC	0.01 0.11	CT	0.11 0.11
0.00 0.10	0.01 0.10	0.10 0.10	0.11 0.10
CA	0.01 0.01	CG	0.11 0.01
0.00 0.00	0.01 0.00	0.10 0.00	0.11 0.00

Figure 1.9: CGR and Markov chains. It is possible to estimate transition probabilities with CGR by counting the points in each sub-quadrant. As an example, by extracting all di-nucleotides counts ($\#\mathsf{CA}, \#\mathsf{CC}, \#\mathsf{CG}, \#\mathsf{CT}$) one can calculate the transition probabilities $p(s_i|\mathsf{C}) = \#\mathsf{C}s_i / \sum_j \#\mathsf{C}s_j$, $s_{i,j} \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. Also shown the addresses of each sub-quadrant represented as binary numbers – the first two decimal bits label each subset and are related to the specific substring (Eq. 1.8).

CGR for higher-order alphabets

More recently a generalization of CGR for higher-order alphabets was defined, in a more natural way, by extending the original CGR square to an hypercube (Almeida and Vinga, 2002). The dimension d of the hypercube will depend on the length of the alphabet $|\mathcal{A}|$ as $d = \log_2 |\mathcal{A}|$. This procedure will be explored in Chapter 3, which transcribes the cited paper.

1.5 Entropy and information theory

The concept of entropy[12] was first introduced in thermodynamics of gases, relating heat and temperature in reversible processes. It was later applied in the modelling of communication systems in the seminal paper by Claude Shannon (1948), which founded the field of information theory (IT). Later the relation between IT and probability and statistics emerged, connecting the concepts of entropy, expected values, mutual information and independence between random variables (Khinchin, 1957; Kullback, 1968).

Entropy concepts are in the base of the work developed on Chapter 5, where a new measure of DNA entropy is proposed. Its introductory section on page 101 contains additional material and relevant historical background.

[12] From Greek *entropé*, in transformation.

1.5.1 Shannon's entropy function

Definitions

Entropy is a measure of the uncertainty associated with a probabilistic experiment. For a discrete random variable X taking values in $\{x_1, x_2, \ldots, x_M\}$ with probabilities $\{p_1, p_2, \ldots, p_M\}$, simplifying as $P(X = x_i) = p_i$, the *Shannon's entropy* H^{Sh} of this experiment is given by Eq. 1.12:

$$H^{\text{Sh}}(X) = H(p_1, \ldots, p_M) = -\sum_{i=1}^{M} p_i \log_2 p_i \tag{1.12}$$

The convention used in this formulation states that $0 \log_2 0 = 0$, justified by continuity since $\lim_{x \to 0} x \log_2 x = 0$. It is also noteworthy that the entropy is a functional[13] of the distribution of X, not depending on the actual values taken by the random variable X.

Shannon's entropy formulation can be interpreted as the minimum number of binary-YES/NO questions necessary in 'average' to determine the output of one observation of X. For example, when tossing a fair coin, the Shannon's entropy is $H^{\text{Sh}}(0.5, 0.5) = 1$ bit. This formulation can also be interpreted in terms of expected values, i.e., $H^{\text{Sh}}(X) = E_p[-\log_2 p(X)]$.

The Shannon's entropy is a non-negative quantity, $H^{\text{Sh}}(X) \geq 0$. It can be shown that $H^{\text{Sh}}(p_1, \ldots, p_M) \leq \log_2 M$ with equality if and only if all $p_i = 1/M$ (Ash, 1990), which means that the situation with the most uncertainty or with the highest entropy occurs when all possibilities are equally likely, thus ascertaining a maximum value for $H^{\text{Sh}}(X)$.

In the original Shannon's formulation, the entropy was expressed in bits as the logarithm base was 2. It is straightforward to change base, since $\log_a b = \log_a x \cdot \log_x b$, so if using natural logarithms $\ln x \equiv \log_e x$, we will obtain $\log_a x = \ln x / \ln a$. Therefore, in the following presentation, the natural base $\ln \equiv \log_e$ will be used, unless otherwise specified.

Axiomatic approach

It can be shown that H^{Sh} is the only function that satisfies the following four axioms (Ash, 1990, chap. 1), where $f(M) = H(1/M, \ldots, 1/M)$ represents the entropy of M equally likely outcomes $p_i = 1/M$.

1. $H(1/M, \ldots, 1/M)$ is a monotonically increasing function of $M \in \mathbb{N}$.

   The entropy should increase with the number of possible equiprobable states;

2. $f(ML) = f(M) + f(L)$

   For independent variables X and Y with possible M and L states, the entropy of the joint experiment with ML equally likely outcomes is equal to the sum of the entropies of the individual experiments;

[13] A functional is a real-valued function on a vector space V, usually of functions.

3. $H(p_1, p_2, \ldots, p_M) = H(p_1 + \cdots + p_r, p_{r+1} + \cdots + p_M) + (p_1 + \cdots + p_r) \cdot H\left(\frac{p_1}{\sum_{i=1}^{r} p_i}, \ldots, \frac{p_r}{\sum_{i=1}^{r} p_i}\right) + (p_{r+1} + \cdots + p_M) \cdot H\left(\frac{p_{r+1}}{\sum_{i=r+1}^{M} p_i}, \ldots, \frac{p_M}{\sum_{i=r+1}^{M} p_i}\right)$
   This grouping axiom permits the deduction of compound experiments;

4. $H(p, 1-p)$ is a continuous function of p.

This result shows that it is possible to deduce Shannon's formulation through an axiomatic approach. All these requirements were initially formulated in the definition of an uncertainty measure.

Other entropy-related measures

Other important notions related to the entropy definition include joint, conditional and relative entropy of two discrete random variables X and Y, with joint probability function $p(x_i, y_j) = P(X = x_i, Y = y_j) = p_{ij}$, $i = 1, \ldots, M$ and $j = 1, \ldots, L$. These measures further deepen the former definition and extended it to the multivariate case, thus permitting the application of new techniques to distribution function comparison.

The *joint entropy* $H(X, Y)$ of X and Y is a natural extension of the former formula and is defined by:

$$H(X, Y) = -\sum_{i=1}^{M} \sum_{j=1}^{L} p_{ij} \ln p_{ij} \tag{1.13}$$

The *conditional entropy* of Y given X is a measure of the average uncertainty of Y *after* the observation of X. It is calculated by:

$$\begin{aligned} H(Y|X) &= \sum_{i=1}^{M} p_i \cdot H(Y|X = x_i) \\ &= \sum_{i=1}^{M} p_i \sum_{j=1}^{L} p(y_i|x_i) \ln p(y_i|x_i) \\ &= -\sum_{i=1}^{M} \sum_{j=1}^{L} p_{ij} \ln p(y_i|x_i) \end{aligned} \tag{1.14}$$

The *relative entropy* or *Kullback-Leibler discrepancy* of the probability mass function $p(X)$ with respect to the mass function $q(X)$ is defined by:

$$D(p||q) = \sum_{i=1}^{M} p_i \ln \frac{p_i}{q_i} \tag{1.15}$$

In the above definition the same convention is used, that $0 \ln \frac{0}{q} = 0$ and $p \ln \frac{p}{0} = \infty$, justifiable by continuity. The relative entropy measures the dissimilarity between two distributions, since $D(p||q)$ is always non-negative and $D(p||q) = 0 \Leftrightarrow p = q$. Although it is not a metric, lacking the symmetrical and the triangle inequality properties (Eq. 1.2), the KL discrepancy has been successfully used in several applications.

The *mutual information* between two random variables X and Y – or the information conveyed about X by Y – is defined as:

$$I(X,Y) = \sum_{i=1}^{M}\sum_{j=1}^{L} p(x_i,y_j)\ln\frac{p(x_i,y_j)}{p(x_i)\cdot p(y_j)} \tag{1.16}$$

Mutual information is a special case of the relative entropy, since $I(X,Y) = D\left(p(X,Y)||p(X)\cdot p(Y)\right)$. Following the properties of $D(p||q)$, the mutual information is 0 if and only if $p(X,Y) = p(X)\cdot p(Y)$, which is the definition of independence between variables X and Y. Therefore, $I(X,Y)$ is measuring the 'dissimilarity' between those variables as assessed by their 'dependence'. Also noteworthy is the relationship between entropy concepts and probability theory already envisaged. One example is given by the alternative definition of independence of events X and Y, specified in terms of conditional probabilities $P(X|Y) = P(X)$, which is very similar to entropy concepts $H(X) = H(X|Y)$.

Some important results relating these measures and proven elsewhere (Ash, 1990; Cover and Thomas, 1991) are summarized below.

1. $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ – chain rule;

2. $H(Y|X) \leq H(Y)$ with equality if and only if X and Y are independent – conditioning reduces entropy;

3. $I(X,Y) = I(Y,X) = H(X) - H(X|Y)$ – if the variables are independent, the mutual information will be zero;

4. $I(X,Y) = D\left(p(X,Y)||p(X)\cdot p(Y)\right)$ mutual information is the relative entropy between the joint distribution and the product of the marginals.

1.5.2 Rényi's entropy generalization

The Rényi formulation appeared has a generalization of the Shannon's measures (Rényi, 1961; Rényi, 1966). In this section the major definitions are presented along with some important properties of this quantity. It will be applied also to continuous probability functions, as a natural extension to the discrete case. Rényi quadratic entropies are the base theoretical framework for the applications presented in Chapter 5, where additional background information can also be found.

Definitions

The Rényi entropy of order $\alpha \geq 0$, $\alpha \neq 1$, H_α is defined both for discrete $p(x)$ and continuous $f(x)$ probability functions and is given by Eq. (1.17). In the following sections, both the continuous and discrete cases are presented to provide a comprehensive and easily accessible listing of important definition and

properties used in most applications.

$$H_\alpha = \frac{1}{1-\alpha} \ln \sum_i p_i^\alpha$$
$$H_\alpha = \frac{1}{1-\alpha} \ln \int f(x)^\alpha dx \tag{1.17}$$

The parameter α here introduced weights each probability function value. When $\alpha \to 0$ the limit of H_α is the logarithm of the support set volume. When $\alpha \to +\infty$ this measure weights more and more the maximum values of p or f and the $H_{+\infty} = \ln(\max_x f(x))$. When $\alpha \to 1$ the limit of Rényi entropy is Shannon's measure, as shown in the next section.

As an example, Fig. 1.10 shows the behavior of the Rényi entropy of a two-state discrete model with probabilities p and $1-p$. $H(X)$ is a concave[14] function of p for $0 < \alpha < 2$.

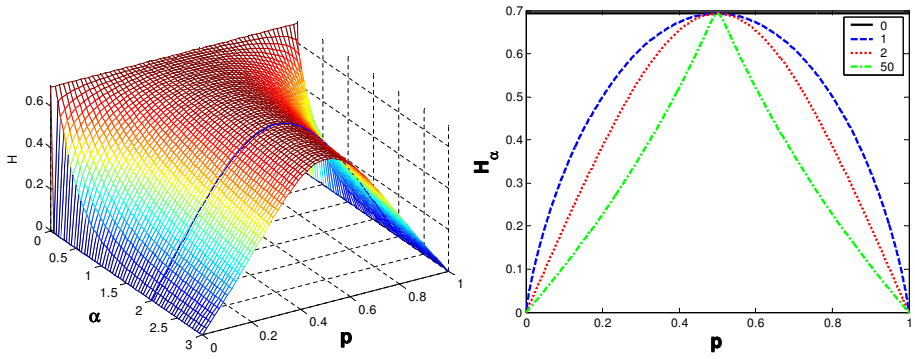


Figure 1.10: Rényi entropy of a two state probability distribution. a) Rényi entropy $H = H(\alpha, p) = \frac{1}{1-\alpha} \ln(p^\alpha + (1-p)^\alpha)$. b) Rényi entropy for $\alpha = 0, 1, 2, 50$ as a function of p. The maximum entropy is obtained when $\alpha = 0$ or $p = 0.5$ and is equal to the logarithm of the number of states $H_{\max} = \ln 2 \approx 0.69$.

The maximum entropy or uncertainty is attained when both states have the same probability $p = 1 - p = 0.5$.

Shannon and Rényi relation

Shannon's entropy is a special case of Rényi's when $\alpha \to 1$. This can be shown using l'Hôpital's Rule :

$$\begin{aligned}
\lim_{\alpha\to 1} H_\alpha(X) &= \lim_{\alpha\to 1} \frac{1}{1-\alpha} \ln \sum_i p_i(x)^\alpha \overset{\left(\frac{0}{0}\right)}{=} \\
&= \lim_{\alpha\to 1} \frac{\frac{\partial}{\partial\alpha} \sum_i p_i(x)^\alpha}{\sum_i p_i(x)^\alpha} \\
&= -\lim_{\alpha\to 1} \frac{\sum_i p_i(x)^\alpha \ln p_i(x)}{\sum_i p_i(x)^\alpha}
\end{aligned} \tag{1.18}$$

[14] A function f is concave if for any two points x and y and $\lambda \in (0,1)$, $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y))$. $f''(x) \leq 0$ if the second derivative exists.

$$
\begin{aligned}
&= -\frac{\sum_i p_i(x) \ln p_i(x)}{\sum_i p_i(x)} \\
&= -\sum_i p_i(x) \ln p_i(x) \\
&= H^{\mathrm{Sh}}(X)
\end{aligned}
$$

An analogous deduction can be made for the continuous case.

Other generalizations

Another important definition related to Rényi formulation redefines the relative entropy between two distributions as a dissimilarity measure. Therefore, *Rényi relative entropy* between p and q is given by:

$$
\begin{aligned}
D_\alpha(p||q) &= \frac{1}{\alpha - 1} \ln \sum_k p_k^\alpha q_k^{1-\alpha} \\
D_\alpha(p||q) &= \frac{1}{\alpha - 1} \ln \int p(x)^\alpha q(x)^{1-\alpha} dx
\end{aligned} \tag{1.19}
$$

Some particular cases are obtained for $\alpha = 1/2$, called the Bhattacharya distance $D_{1/2}(p||q) = -\ln \int \sqrt{f(x)g(x)}dx$. Analogously, and using the same deduction of Eq. (1.18), the limit of this measure when $\alpha \to 1$ is the relative entropy defined in Eq. (1.15).

1.6 Thesis outline

Each of the following chapters constitutes the transcription of one published paper, to which additional material might also be added, opportunely indicated in their respective introductory section. The rationale for this design is justified in the preface to the thesis.

A simplified scheme of this work is shown in Fig. 1.11, with suggested reading lines.

Chapter 2 – Alignment-free sequence comparison – a review – is a literature review of alignment-free sequence comparison methods and introduces the motivation for all the subsequent work (Vinga and Almeida, 2003). This overview of methods not based in alignment provides a background section on words in sequences and imparts several definitions that will be used later in this work. An effort was also made to uniform the nomenclature, consequently altering the original formulations, in order to establish a common ground for further methods development. This bibliographic review can also be seen as a continuation of the present introductory chapter, as it presents more background information.

Chapter 3 – Universal sequence map (USM) of arbitrary discrete sequences, preceded chronologically the review (Almeida and Vinga, 2002). In this work, a natural extension of CGR maps is proposed, allowing the representation of higher-order alphabet sequences. Several properties are explored, namely the study of the backward sequence coordinates and measures of dissimilarity between symbols. This formulation will be pivotal to the calculations of DNA sequences entropies, on Chapter 5.

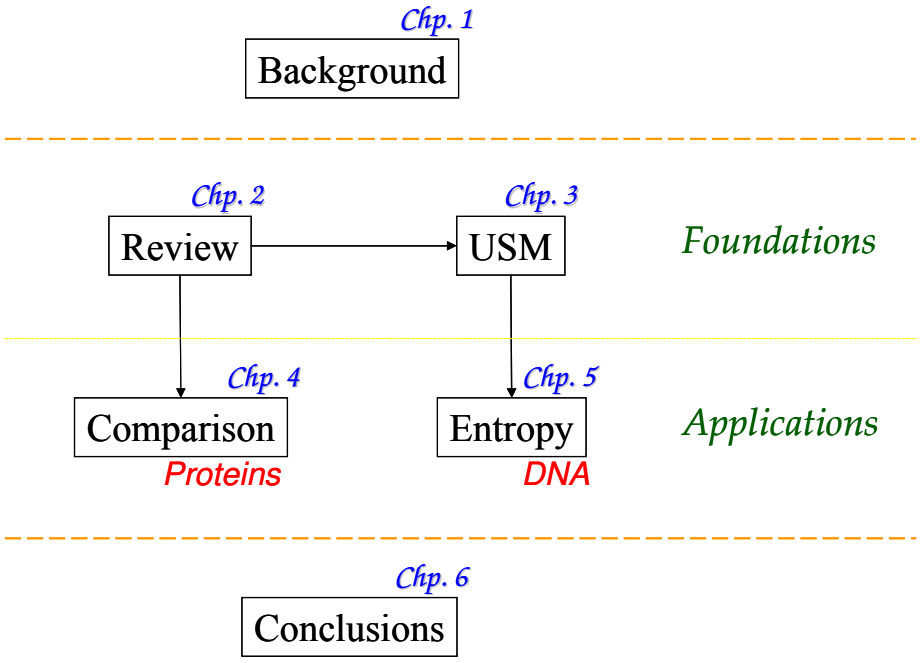


Figure 1.11: Thesis outline scheme. Suggested reading lines.

These two chapters constitute the theoretical foundations of the thesis, defining two distinct types of vector mappings that belong to the alignment-free category in sequence analysis. The first one, extensively reviewed, maps the sequences onto their L-tuple frequency vector, accounting for the relative abundance of L-length substrings. The second method is based on iterated function systems and maps symbols through an iterative algorithm. This mapping has context-based properties since it is possible to recover all the sequence from the mapping of just one symbol.

The next two chapters are devoted to applications of these methods to biological sequences, namely the classification of proteins and the estimation of DNA entropy.

Chapter 4 – Comparative evaluation of word composition distances for the recognition of SCOP relationships – qualitatively analyzes some of the metrics described before and proposes a new metric, W-metric, that bridges between alignment-free methodologies (based on tuple counts) and alignment-based algorithms (Vinga et al., 2004). This intersection is achieved by conjugating scoring mutational matrices with L-tuple based information. An extensive comparative evaluation of the dissimilarity measures previously reviewed is accomplished by confronting the classification results of protein secondary structure. For this assessment, ROC curves theory and the Structural Classification of Proteins (SCOP) database, a gold standard for structure prediction, are used as major accuracy evaluation techniques.

Chapter 5 – Rényi continuous entropy of DNA sequences – presents a CGR/USM-driven entropy definition, based on Rényi formalism, which con-

stitutes a novel application of iterative maps (Vinga and Almeida, 2004). The novelty consist on spanning the parameter space continuously using the Parzen window density estimation method with gaussian kernels. Some examples of testing are also presented, with the application to artificial and real DNA. Additionally, theoretical properties of the measure are deduced, namely its asymptotical behavior. Furthermore, Monte Carlo simulations are also performed to estimate the variability of this quantity.

These two chapters represent the application part of the work outlined in the figure, with both the study of proteins and DNA, in two specific distinct problems. In the protein study – Fig. 1.11 left side – all pairwise comparisons are made in order to classify the sequences into categories, therefore only the relations between the sequences are important. On the other hand, for the DNA example – Fig. 1.11 right side – the interest is to analyze each sequence separately, measuring their global degree of entropy or uncertainty, with no connection with comparison techniques. These applications confirm the flexibility of these methods, in terms of the objects to which they can be applied and the target results intended.

The two suggested reading lines are now evident: from the review to the proteins classification application, and from the review and through the generalization of iterative maps to the analysis of individual DNA sequences by Rényi entropy-based methods.

Chapter 6 – Final discussion – presents an overall discussion of all the themes treated, briefly recalls the main thesis' achievements and suggests future non-exploited paths in this particular field. This concluding chapter also describes open problems and epistemological issues in bioinformatics and how the recent development and revision of strong paradigms and dogmas might change the future of this field.

1.7 References

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., and Fletcher, M. (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics*, 17(5):429–437. 20

Almeida, J. S. and Vinga, S. (2002). Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*, 3(1):6. 15, 18, 22, 27

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215:403–410. 14

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402. 14

Ash, R. B. (1990). *Information Theory*. Dover Publications, New York. 23, 25

Barnsley, M. F. (1998). *Fractals Everywhere*. Academic Press, Boston. 15

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). Genbank: update. *Nucleic Acids Res*, 32 Database issue:D23–6. 2

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York. 25

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press. 11

Edgar, G. A. (1990). *Measure, topology, and fractal geometry*. Undergraduate texts in mathematics. Springer-Verlag, New York. 15, 16

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press. 11

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res*, 18(8):2163–2170. 17

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. Dover Publications, New York. 22

Kulkarni, V. G. (1995). *Modeling and Analysis of Stochastic Systems*. Chapman & Hall, New York. 20

Kullback, S. (1968). *Information theory and statistics*. Dover Publications, New York. 22

Lodish, H. F., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, L., and Darnell, J. (2004). *Molecular cell biology*. W.H. Freeman and Company, New York, 5th edition. 8, 10

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48:443–453. 12

Nester, E. W., Anderson, D. G., Roberts, Jr., C. E., Pearsall, N. N., and Nester, M. T. (2004). *Microbiology: a human perspective*. McGraw-Hill, Boston, 4th edition. 7

Ouzounis, C. A. and Valencia, A. (2003). Early bioinformatics: the birth of a discipline–a personal view. *Bioinformatics*, 19(17):2176–90. 3

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, 183:63–98. 14

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 85(8):2444–8. 14

Rényi, A. (1961). On measures of entropy and information. In *Proc. of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, volume 1, pages 547–561. University of California Press. 25

Rényi, A. (1966). Introduction a la théorie de l'information. In *Calcul des probabilités*. Dunod, Paris. 25

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656. 22

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147:195–197. 14

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80. 14

Tino, P. and Dorffner, G. (2001). Predicting the future of discrete sequences from fractal representations of the past. *Machine Learning*, 45:187–217. 21

Tortora, G., Funke, B., and Case, C. (2004). *Microbiology: an introduction*. Pearson Education, Inc., 8th edition. 5

Trifonov, E. N. (2000). Earliest pages of bioinformatics. *Bioinformatics*, 16(1):5–9. 3

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison–a review. *Bioinformatics*, 19(4):513–523. 11, 14, 27

Vinga, S. and Almeida, J. S. (2004). Rényi continuous entropy of DNA sequences. *J Theor Biol*, 231(3):377–388. 15, 29

Vinga, S., Gouveia-Oliveira, R., and Almeida, J. S. (2004). Comparative evaluation of word composition distances for the recognition of SCOP relationships. *Bioinformatics*, 20(2):206–215. 14, 15, 28

Weisstein, E. W. (2004). "MathWorld–A Wolfram Web Resource." Accessed: 30 Nov. 2004. <*http://mathworld.wolfram.com*>.

Chapter 2

Alignment-free sequence comparison – a review

Published in: Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison – a review. Bioinformatics *19:4, 513–523.*

Supplementary material added: Non-aligned sequence comparison (NASC) toolbox description along with three classification applications to DNA, Proteins and Natural Languages texts – extended section 2.4. All MATLAB functions described are available online.

Motivation Genetic recombination and, in particular, genetic shuffling are at odds with sequence comparison by alignment, which assumes conservation of contiguity between homologous segments. A variety of theoretical foundations are being used to derive alignment free-methods that overcome this limitation. The formulation of alternative metrics for dissimilarity between sequences and their algorithmic implementations are reviewed.

Results The overwhelming majority of work on alignment-free sequence has taken place in the past two decades, with most reports published in the past 5 years. Two main categories of methods have been proposed - methods based on word (oligomer) frequency, and methods that do not require resolving the sequence with fixed word length segments. The first category is based on the statistics of word frequency, on the distances defined in a Cartesian space defined by the frequency vectors, and on the information content of frequency distribution. The second category includes the use of Kolmogorov complexity and chaos theory. Despite their low visibility, alignment-free metrics are in fact already widely used as pre-selection filters for alignment-based querying of large applications. Recent work is furthering their usage as a scale-independent methodology that is capable of recognizing homology when loss of contiguity is beyond the possibility of alignment.

Availability Most of the alignment-free algorithms reviewed were implemented in MATLAB code and are available at
http://bioinformatics.musc.edu/NASC.

2.1 Introduction

Sequence analysis is a discipline that grew enormously in recent years in response to the overwhelming burst in data generated by molecular biology initiatives. This tendency will probably continue as new challenges emerge from its quantity and increasingly integrative nature (Fuchs, 2002; Reichhardt, 1999). Although initially the algorithms were mostly borrowed from string processing computer science methodologies (Gusfield, 1997), in a second stage biological sequence analysis quickly incorporated additional concepts and algorithms from computational statistics, such as stochastic modelling of sequences using hidden Markov models and other Bayesian theory methods for hypothesis testing and parameter estimation. Both foundations carry a bias, very clear in present days, that views biological molecules as being linear sequences of discrete units similar to linguistic representations, in spite of their physical nature as a 3D structure and the dynamic nature of molecular evolution. The alignment approach overlooks well-documented long-range interactions and general fluidity resulting from recombination with shuffling of conserved segments without loss of function (Zhang et al., 2002; Lynch, 2002). On the other hand, assuming conservation of contiguity allows the employment of a large set of well-developed effective computational procedures. Accordingly, the use of alignment based pairwise sequence comparison emerges in many bioinformatic applications associated with querying a sequence databases with a template, where sequence similarity is used to infer similar structure or function. Moreover, sequence divergence, leading to dissimilarity between homologous sequences, is intrinsically hard to solve as the evolutionary process takes place at different scales simultaneously (Attwood, 2000; Pearson, 2000).

The difficulty in defining a metric for sequence dissimilarity is also apparent in the analysis of natural languages texts (Searls, 2001). The quantification of similarity between texts is not unique and unambiguous, depending strongly on the relative importance assigned to individual particles, letters, words, phonemes, and grammar and even to the overall context of its occurrence. The overwhelming majority of biological sequence comparison methods rely on first aligning reference homologous sequences and deriving a score for the alignment of individual units, typically the logarithm of the odds ratio. This score is then used to optimize the alignment of new sequences. Consequently, sequence dissimilarity is reduced to the comparison between candidate alignments and reference alignment of well-studied sequences, a heuristic solution for a fundamental problem for which effective solution remains open. Although alignment methods are not reviewed here, comprehensive reviews abound (Durbin et al., 1998; Waterman, 1995), a very brief overview of the context of its present wide use is warranted. There are two basic aspects to consider – the alignment itself and the scoring used to produce it. Optimal sequence alignment algorithms are implemented using dynamic programming, ultimately a regression technique that identifies optimal alignment by maximizing the score of the path that produces it. Several algorithms have long been identified that target specific goals such as global alignment, local alignment, with or without overlapping (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Gotoh, 1982). Although

the algorithmic solutions appear satisfactory, the computational load escalates as a power function of the length of the sequences (exponent 2 for un-gapped alignment and somewhat higher for the best gapped algorithms) making its use for searching large databases unfeasible. Subsequently, a few heuristic approaches were proposed, mostly based on the recognition of alignment "seeds", with BLAST (Altschul et al., 1990, 1997) and FASTA (Pearson, 1990; Pearson and Lipman, 1988) being the most ubiquitous applications. The second critical consideration in this reference to alignment methods is the scoring of pairwise unit alignments. A wide range of scoring systems has been proposed such as aminoacid substitution scoring matrices PAM (Dayhoff et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992) for protein alignment. This heuristic solution reflects methodological incompleteness in the approach to sequence divergence, and also reflects assumption of conservation of contiguity between homologous segments. It is interesting to note that no scoring schemes in use will consider increasing its memory length, e.g. scoring alignment of individual oligomers rather than of individual units, equivalent to using higher order Markov model scores.

The more immediate limitations of alignment based sequence analysis are consistently restated in all the reports reviewed below. Another difficulty, not often discussed, is that heuristic solutions make it harder to assess the statistical relevance of the resulting scores, which compromises, for example, the establishment of confidence intervals for homology. Nevertheless, the distribution of the maximum score obtained under the null hypothesis (non correlated sequences) was deducted recently for gapped alignments (Siegmund and Yakir, 2000; Storey and Siegmund, 2001) providing a long waited reinforcement of the theoretical foundations of scoring methods.

This report reviews published concepts and the corresponding algorithms for alignment-free comparison of biological sequences. In spite of the present surge in interest on alignment-free sequence comparison methods, there has never been, to our knowledge, any collective review of published work. However, classification, clustering or grouping techniques are not included in this overview. In cluster analysis the basic input is the cross-tabulation of dissimilarity which is then object of agglomeration, for which there is extensive literature and widely available implementation in standard statistical packages. (For a comprehensive introduction to cluster analysis and classification see (Everitt et al., 2001; Gordon, 1999).) This review is confined to the measure of sequence dissimilarity itself.

2.2 Background

The variety of disciplines involved in development of biological sequence analysis often brings together conflicting nomenclatures and conceptual frameworks. Therefore, for the convenience of the reader, some useful concepts and notation in vectors and metric spaces, information theory and word statistics are briefly recalled. References to comprehensive presentations of those fields are also included for further depth.

2.2.1 Words in sequences

A sequence, X, of length n, is defined as a linear succession of n symbols from a finite alphabet, $\mathcal{A}$, of length r.

A segment of L symbols, with $L \leq n$, is designated an L-tuple (in some references is also defined as L-word or L-plet). The set $\mathcal{W}_L$ consists of all possible L-tuples that can be extracted from sequence X and has K elements (Eq. 2.1)

$$\begin{aligned} \mathcal{W}_L &= \{w_{L,1}, w_{L,2}, \ldots, w_{L,K}\} \\ K &= r^L \end{aligned} \tag{2.1}$$

The identification of L-tuples in the sequence X can then be object of counting occurrences with overlapping (Eq. 2.2). Computationally, the counting is usually performed by taking a sliding window L-wide that is run through the sequence, from position 1 to $n - L + 1$.

$$c_L^X = (c_{L,1}^X, c_{L,2}^X, \ldots, c_{L,K}^X) \tag{2.2}$$

Similarly, one can then calculate word frequencies, f_L^X , to estimate the probability, $p_{L,i}^X$, of finding a specific word $w_{L,i}^X$, collectively defining a vector of word or L-tuple probabilities (Eq. 2.3).

$$p_L^X = (p_{L,1}^X, p_{L,2}^X, \ldots, p_{L,K}^X) \tag{2.3}$$

The vector of frequencies f_L^X is obtained as the relative abundance of each word (Eq. 2.4)

$$f_L^X = \frac{c_L^X}{\sum_{j=1}^K c_{L,j}^X} \Leftrightarrow f_{L,i}^X = \frac{c_{L,i}^X}{n - L + 1} \tag{2.4}$$

For example for DNA sequences, $\mathcal{A} = \{\mathsf{A}, \mathsf{T}, \mathsf{C}, \mathsf{G}\}$, $r = 4$, a three letter word, $L = 3$, could be $w_3 = \mathsf{ATC}$. For the sequence $X = \mathsf{ATATAC}$, where $n = 6$, the vector p_3^X is estimated by the relative frequencies of all trinucleotides. The frequencies, determined by sliding a 3 letter window $n - L + 1 = 4$ times would be:

$$\begin{aligned} \mathcal{W}_K &= \{\mathsf{ATA}, \mathsf{TAT}, \mathsf{TAC}, \mathsf{AAA}, \ldots\} \\ c_3^X &= (2, 1, 1, 0, \ldots) \\ f_3^X &= (0.5, 0.25, 0.25, 0, \ldots) \end{aligned}$$

The vectors c_3^X and f_3^X have length $K = 4^3 = 64$ and the zero coordinates correspond to missing words in X, in this case absent trinucleotides.

2.2.2 Distance between sequences

A distance function $d(X, Y)$ is a function that assigns a real number to every pair X and Y belonging to a given set, in this application will be the set of all

possible sequences. In order for $d(X,Y)$ to be a metric distance (Strang, 1988) the three properties in Eq. 2.5 have to be observed.

$$\begin{aligned} \text{Positivity} &: d(X,Y) \geq 0 \quad \text{and} \quad d(X,Y) = 0 \Leftrightarrow X = Y \\ \text{Symmetry} &: d(X,Y) = d(Y,X) \\ \text{Triangle inequality} &: d(X,Y) + d(Y,Z) \geq d(X,Z) \end{aligned} \tag{2.5}$$

Most of the distance functions reviewed below are computed in the spaces defined by the vectors of word counts and word frequencies. For a comprehensive introductory study of linear algebra and vector spaces see (Strang, 1988) and for an introduction to matrix analysis (Schott, 1997) is recommended.

2.2.3 Word statistics

The statistical and probabilistic properties of words in sequences were recently systematized and reviewed (Reinert et al., 2000), with emphasis on the deductions of exact distributions and the evaluation of its asymptotic approximations. The problems addressed in that report included finding formulae for counts expectation, variances and also covariances between frequencies of two words, namely the distribution of p_L^X and the determination of its moments. These issues are fundamental to assess the statistical significance of dissimilarity results based on frequencies of words. The period or overlap capability deserves special mention here, as it will be of importance for the reviewing, below, of metrics based on the Mahalanobis distance. It indicates to what extent the prefix and the suffix of a word are equal, i.e., if the word beginning is the same as the ending (Gentleman and Mullin, 1989). This property is fundamental to the correct deduction of the covariances of p_L^X , as words that share motifs are more likely to co-occur. The modeling of the resulting word statistics is often approached within the framework of the theory of stochastic processes, namely Markov chains and renewal theory (Gentleman and Mullin, 1989; Régnier, 1998; Reinert et al., 2000; Gentleman and Mullin, 1989; Régnier, 1998; Reinert et al., 2000; Waterman, 1995) and will not be reviewed here.

2.2.4 Information theory

Information theory was originated in the classical paper of Claude Shannon in 1948 (Shannon, 1948) to quantify the capability of transmitting data over a channel. Some years later, Solomon Kullback formalized it as a branch of statistical theory (Kullback, 1968) and gave rigorous mathematical proofs of theorems previously introduced. The main concept behind information theory is the notion of entropy or uncertainty. One defines the entropy of a random variable based on the probabilities of all the outcomes. The definition will be subsequently applied to sequences, where the random variable represents an L-tuple. The entropy H of L-tuples, $\mathcal{W}_L$, is calculated from the probability of the individual words in sequence X (Eq. 2.6).

$$H(\mathcal{W}_L^X) = -\sum_{i=1}^{K} p_{L,i}^X \log_2(p_{L,i}^X) \tag{2.6}$$

This general definition is valid for any word length resolution, L, including the more common determination of uncertainty associated to the distribution of individual symbols, e.g. by using $L = 1$. It was subsequently shown that this is the only function that satisfies some logically required axioms for the quantification of uncertainty (Ash, 1990), such as additively of entropies for joint probability spaces, the fact that $H(\mathcal{W})$ is maximal when all the K possible words are equiprobable, $H(\mathcal{W}) = \log_2(K)$, and it is minimal when $p_{L,i}^X = 1$ for some i-word – knowing the outcome should make uncertainty equal to zero. $H(\mathcal{W})$ is also an increasing function of K equiprobable spaces, i.e., it will be higher if the number of possible words increase. Comprehensive presentations of this matter and respective applications abound, such as (Cover and Thomas, 1991). For the studies reviewed below it is useful to detail the Kullback-Leibler (KL) discrepancy, measuring relative entropy between two discrete probability distributions p and q, detailed in Eq. 2.7.

$$KL(p, q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \tag{2.7}$$

However, it is noteworthy that the KL discrepancy is not a metric distance because it only satisfies positivity but not symmetry nor triangular inequality (Eq. 2.5).

2.3 Alignment-free sequence comparison

The proposition of alignment-free methods to compare biological sequences is a very recent endeavor, with the earliest systematic publications being less than 2 decades old (Blaisdell, 1986). Although the pace of work in this area is increasing sharply, the total number of published reports proposing or using alignment-free metrics is relatively small, still under the one hundred mark. Moreover, the past decade contains the overwhelming majority of reports and judging by those published in the past year, the trend is being maintained. Two main categories of proposed methods can be recognized in the literature reviewed – methods based on word frequency, and those that do not require resolving the sequence with fixed word-length segments. The first group includes procedures based on metrics defined in coordinate space of word-count vectors, such as the Euclidean distance and relative entropy of frequency distributions. On the contrary, the second category corresponds to techniques that are independent from the resolution of the sequence, i.e., they do not involve counting segments of fixed length. They include the use of Kolmogorov complexity theory and scale-independent representation of sequences by iterative maps. These two categories of methods have distinct theoretical lineages and unequal amount and variety of techniques explored in the published reports, far fewer for the latter.

2.3.1 Methods based on word frequencies

All methods described in this section start with the mapping of sequences to vectors defined by the counts of each L-tuple. This straightforward approach was the first attempt to transform a sequence into an object for which linear algebra and statistical theory had useful analytical tools already available. The vectors obtained represent the original sequence with a fixed resolution L, that of the word length considered. The basic rationale for sequence comparison is that similar sequences will share word composition to some extent, which is then quantified by a variety of techniques. This is, in a way, an extension of the widespread use of difference in GC content as a measure of sequence dissimilarity. It is noteworthy that the methods described here, although alignment-free, are still length dependent in the sense that the comparisons are made for fixed word length. This could even be viewed as a weak departure from the idea of alignment since sharing L-tuples is equivalent to recognizing an alignment between identical segments. However, a variety of methods have been proposed to derive combined distance metrics that contain information about all resolutions, in order to achieve complete independence from the contiguity of conserved segments.

Euclidean distance

The first published report systematizing the use of L-tuple counts for sequence comparison dates from 1986 (Blaisdell, 1986). In this work, the author presents a new measure of dissimilarity between sequences modeled as Markov chains. The difference between two sequences was quantified by the square Euclidean distance between their transition matrices. In spite of its conceptual simplicity, this method was shown to be an effective alternative to alignment methods. The fact that a transition matrix of a Markov chain can be identified with the frequency of all L-tuples lead the author to propose other quantifications of sequence similarity, such as the use of Chi-Square tests to assess the statistical significance of a specific comparison (Blaisdell, 1986). It was further shown in this pioneering report that the approach enabled the measure of dissimilarity between sequences that are too different to be amenable to alignment, even if they still have recognizable similarity. The fact that, when alignment is possible the two methods agree, provided further support for the adoption of the more generally applicable alignment-free alternative. For each resolution or word-length L, the squared Euclidean distance between sequences X and Y is determined by Eq. 2.8, where $c_L^X = (c_{L,1}^X, \ldots, c_{L,K}^X)$ and $c_L^Y = (c_{L,1}^Y, \ldots, c_{L,K}^Y)$ are vectors representing word counts for those sequences and K is the number of different L-tuples possible for that L-length:

$$\begin{aligned} d_L^E(X,Y) &= (c_L^X - c_L^Y)^T \cdot (c_L^X - c_L^Y) \\ &= \sum_{i=1}^{K} (c_{L,i}^X - c_{L,i}^Y)^2 \end{aligned} \tag{2.8}$$

Nevertheless, alignment was still observed in the same report to be more accurate for comparison of sequences with very close similarity. A few years later

the same author formalized the new alignment-free metric and validated its performance by successfully comparing large genomic sequences from organisms with well documented phylogenetic relationships (Blaisdell, 1989b). The dissimilarity values obtained by pairwise sequence comparison was subsequently used to correctly recognize phylogenetic relationships with PHYLIP package (Felsenstein, 1993), corroborating results obtained with 'conventional methods that assume prior correct homologous total alignment of the sequences'. A similar conclusion was reached in a subsequent study (Blaisdell, 1989a) where the dissimilarity values obtained with alignment-free Euclidean distance were observed to be directly proportional to conventional mismatch counts requiring sequence alignment. A subsequent report presented statistical deductions of several characteristic measures (Pevzner, 1992) such as the distance expectation and variance for L-tuple comparison. The same report proposes filtration methods based on a prescreening with these metrics. Accordingly, it is possible to filter out sequences with low similarity, those that do not share similar word composition, in order to speed database search for similar sequences. In that report, the same theoretical endpoint proposed previously (Blaisdell, 1989a) is reached. It is noteworthy that these filtration methods are currently being increasingly explored to optimize database search in the face of exponential growth of the sequence repository (URL, 2002).

The statistical properties of Euclidean type distance for L-tuple frequencies have been documented further in depth eventually leading to the identification of tests for the non-uniformity of the corresponding distribution based on the Π-statistic thereby defined (Zharkikh and Rzhetsky, 1993). This work enabled the comparison of values obtained for different resolutions and also offers the very interesting promise of a formal link to the determination of evolutionary distances backed by a rate of unit substitution that is not affected by shuffling of conserved segments. The same authors also document a relation between L-tuple metric and mismatch count distance, which is the basis for homology estimation by alignment-based methods, thus establishing some comparison between both methods. The validity of those theoretical propositions was accessed in another report with applications to Eubacteria, mitochondria and chloroplasts DNA, including the study of L-tuple frequency homogeneity in coding and non-coding regions (Sitnikova and Zharkikh, 1993). The scale dependency of similarity measures itself, such as how 3-tuple counts depends on 2-tuple counts described in the latter report, is also becoming a recurring theme, albeit reinforced by similar emphasis in the search for unifying scale independent relationships in other areas of Biology (Gisiger, 2001).

Weighted Euclidean distance and efficient computation – The fact that the frequency of different words may have different impact on the standard Euclidean distance between specific words has been explored in the literature to derive weighted measures. The earliest work calculated the weights of individual L-tuples in order to maximize the variance of reference sequences with regard to random sequences (Torney et al., 1990). This approach maximizes the discrimination of reference sequence families. The original implementation, maybe due to its relatively pioneering date of 1990, is curiously based on weighting L-tuple counts rather than frequencies (Eq. 2.9), where ρ_i is the weight assigned to the

i^{th} word. The weighted distances are then combined by summing the weighted count difference at different resolutions, from l to u-tuples.

$$d^2(X,Y) = \sum_{L=l}^{u} \sum_{i=1}^{K} \rho_i (c_{L,i}^X - c_{L,i}^Y)^2 \tag{2.9}$$

This metric was designated as d2 distance and has subsequently been used, in its unweighted form, as a stand-alone high performance sequence comparison technique for database search (Hide et al., 1994). The latter work stands on a category of its own due to its focus on heuristic optimization of the computational implementation. That report in particular was directed to the identification of optimum values for word length L, window size and extent of overlap. For the particular example discussed in that report, search for lipases in a genomic database, an optimal resolution of $L = 8$ was found to achieve results similar to performing the search using FASTA.

The practical use of d2 distance has a published record that continues to present day including the clustering of EST sequences with full-length cDNA data (Burke et al., 1999) and the recent estimation of the number of human genes (Davison and Burke, 2001). The method has proven to be selective, sensitive and amenable to high performance implementation. These properties, combined with the advantages shared by other alignment-free methods of being context-independent, and consequently the fact that homologous sequences that are scrambled or contain insertions and deletions will still yield a small d2 value, has had this measure selected for inclusion in software packages. In particular, d2 clustering was incorporated in the software package STACK (Sequence Tag Alignment and Consensus Knowledgebase), a sequence analysis tool where clustering does not rely on pairwise alignment (Burke et al., 1998; Christoffels et al., 2001; Hide et al., 1997; Miller et al., 1999). Even more recently, this algorithm was optimized by parallelization (Carpenter et al., 2002), furthering their efficient computation, with a visible relevance for the classification of EST sequences.

In general, it is interesting to note that, very recently, filtration methods based on distance between frequencies of words have had their usage greatly increased as procedures to "seed" a conventional alignment, both for DNA sequences (Giladi et al., 2002) and for proteins (Coghlan et al., 2001). Both FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990) rely on seeding for a pre-selection of candidate sequences for alignment. Indeed, preprocessing sequence querying by efficient elimination of non-similar candidates appears to be the path through which alignment-free sequence comparison is gradually being incorporated in widely used bioinformatics applications.

Correlation structure

Once the conversion of sequences into L-tuple frequencies was established, a variety of metric systems were quickly proposed, as described above for Euclidean distances. Within this context, the proposition of metric distances between sequences based on the correlation coefficients was to be expected (Fichant and Gautier, 1987; Gibbs et al., 1971; van Heel, 1991). Indeed, that approach has

since been put to practice to classify proteins based on di-peptide frequencies (Petrilli, 1993). The calculation of the linear correlation coefficient (LCC) between two sequences X and Y, from L-tuple frequencies, f_L^X and f_L^Y, uses the conventional Pearson formalism as detailed in Eq. 2.10.[1]

$$d_L^{LCC}(X,Y) = \frac{K\sum_{i=1}^{K} f_{L,i}^X \cdot f_{L,i}^Y - \sum_{i=1}^{K} f_{L,i}^X \cdot \sum_{i=1}^{K} f_{L,i}^Y}{\left(K\sum_{i=1}^{K}(f_{L,i}^X)^2 - (\sum_{i=1}^{K} f_{L,i}^X)^2\right)^{1/2} \times \left(K\sum_{i=1}^{K}(f_{L,i}^Y)^2 - (\sum_{i=1}^{K} f_{L,i}^Y)^2\right)^{1/2}} \tag{2.10}$$

This can be simply expressed by taking vectors f_L^X and f_L^Y as pairs in $\mathbb{R}^2$, by plotting the K points $(f_{2,i}^X, f_{2,i}^Y)$, and calculating the correlation coefficient R. As noted before for Euclidean distances, the availability of a correlation based, alignment-free, sequence comparison method is of immediate advantageous use for the querying of large sequence databases, and has been applied to protein database searching (Petrilli and Tonukari, 1997). The applied work yielded a number of simplifying conclusions that greatly enhance its practical value, such as the fact that only 25 out of 400 possible dipeptide frequencies were needed to correctly classify protein families (Solovyev and Makarova, 1993).

The way tuples are defined has itself been object of exploration with the goal of identifying spatial correlations between positions differently spaced apart in the sequence (Mironov and Alexandrov, 1988). Although this approach has not been subsequently pursued by other researchers, its original proposition took place in the very early period of development of alignment-free methods and offers a different perspective on the conceptual foundations of this field. The spatial correlation measure is based on the determination of dimeric tuples ($L = 2$) where the first and second positions are separated by a fixed arbitrary number of units. The original report proposed to screen different values for the separation and combine the results in a single correlation measure. The difference between sequences was then developed using the Euclidean distance of the vectors representing the extracted features.

Covariance methods

The methods reviewed above explore the use of Euclidean distances and correlations between L-tuple representations of sequences. This section reviews, instead, distances that take into account the data covariance structure. In this context the use of Mahalanobis distances (Eq. 2.11) and standardized Euclidean distances (Eq. 2.12), play a central role.

[1] Original formulation – can be further simplified using probability properties.

$$\begin{aligned} d_L^M(X,Y) &= (c_L^X - c_L^Y)^T \cdot S^{-1} \cdot (c_L^X - c_L^Y) \\ &= \sum_{i=1}^{K}\sum_{j=1}^{K}(c_{L,i}^X - c_{L,i}^Y) \cdot s_{ij}^{\text{inv}} \cdot (c_{L,j}^X - c_{L,j}^Y) \end{aligned} \tag{2.11}$$

In Eq. 2.11, $S = [sij]$ represents the covariance matrix of L-tuple counts, which inverted is composed of $K \times K$ elements s_{ij}^{inv} . The standard Euclidean distance (Eq. 2.12) forces $\text{cov}(c_i, c_j) = 0$ for $i \neq j$. Therefore, in this distance measure the correlations between different words are ignored and only same word variances are accounted for.

$$\begin{aligned} d_L^{SE} &= (c_L^X - c_L^Y)^T \cdot [\text{diag}(s_{11}, \ldots, s_{KK})]^{-1} \cdot (c_L^X - c_L^Y) \\ &= \sum_{i=1}^{K} \frac{(c_{L,i}^X - c_{L,i}^Y)^2}{s_{ii}} \end{aligned} \tag{2.12}$$

The relevance of this simplification is put into context by noting that the standard Euclidean distance (Eq. 2.12) is reduced to the squared Euclidean distance (Eq. 2.8) if the variance structure is ignored, i.e., if $s_{ii} = 1$, $i = 1, \ldots, K$. Both the Mahalanobis and standard Euclidean distance were first proposed for sequence comparison relatively recently (Wu et al., 1997). In that report the author also proposes to combine different resolutions to obtain a unique distance measure (Eq. 2.13), similarly to the approach followed in the definition of the d2 measure (Eq. 2.9).

$$\begin{aligned} d^{M*} &= \sum_{L=l}^{n} d_L^M \\ d^{SE*} &= \sum_{L=l}^{n} d_L^{SE} \end{aligned} \tag{2.13}$$

It is in the context of these metrics that the measure of overlap capability between words, introduced in the Background section above, is most relevant. Overlap capability indicates periodicity in the word, which leads to higher probability of co-occurrence of words sharing the repeated motifs (Gentleman and Mullin, 1989; Reinert et al., 2000), consequently altering the covariance structure presented.

Some implementation problems arise when calculating Mahalanobis distance: the covariance matrix S has determinant near zero (matrix almost singular) so it is computationally difficult to calculate its inverse. A solution often proposed that was followed to overcome this problem is to use pseudo inverse matrices (Wu et al., 1997). However this is unsatisfactory for word lengths higher than 4, when the computational load becomes too heavy for practical implementation. For this reason and although important from a theoretical point of view, this method was ruled out by the proponent for applications with long alphabets and/or long sequences. Nevertheless, it was shown to be

very efficient when challenged with finding human lipoprotein lipase (LPL) in a database, providing better selectivity and sensibility than previous distances, namely the Euclidean and standard Euclidean measures.

The Mahalanobis based distance was also proposed for protein classification in a report (Solovyev and Makarova, 1993) already approached in the Correlation section. With regard to the Mahalanobis distance, the proponents suggested practical simplifications, namely that only oligopeptides whose frequencies are distinct from random proteins, are used, as these are the most informative and, consequently, the most discriminant data.

Information theory based measures

The methods reviewed above were based on statistical distances between frequency vectors. Instead, the distances reviewed in this section are based on the same L-tuple vectors as above but an information theory based metric is use to quantify the dissimilarity between them. To that effect, the Kullback-Leibler discrepancy, KL (see Background section), was recently proposed (Wu et al., 2001). The KL discrepancy between sequences X and Y, is computed from their L-tuple frequencies (Eq. 2.14).

$$d_L^{KL}(X,Y) = \sum_{i=1}^{K} f_{L,i}^X \cdot \log_2 \left(\frac{f_{L,i}^X}{f_{L,i}^Y} \right) \tag{2.14}$$

To avoid having an infinite $d_L^{KL}(X,Y)$ when $f_{L,i}^Y = 0$, the authors also suggest modifying this formulation (Eq. 2.14) by adding a unit to both terms of the frequency ratio. As with the Mahalanobis distance, this report also proposes an implementation by sliding partially overlapping windows to select the best conserved regions, under the assumption of contiguity discussed above. The KL distance was validated using the human lipoprotein lipase dataset the same authors had previously used to evaluate the use of Mahalanobis distance (Wu et al., 1997). It was concluded that the best performing metric with regard to selectivity and sensitivity was the Mahalanobis distance (Eq. 2.11), followed closely by the standard Euclidean distance (Eq. 2.12) and somewhat further behind by the KL discrepancy (Eq. 2.14). These three distance measures clearly outperformed the conventional Euclidean distance (Eq. 2.8). As regards computational efficiency, the performances are reversed with KL discrepancy (Eq. 2.14) being preferred, followed by the standard Euclidean distance (Eq. 2.13). The Mahalanobis distance, as mentioned above, has a hefty computational cost associated to the calculation of the inverse covariance matrices S^{-1} (Eq. 2.11).

Angle metrics

Very recently, (Stuart et al., 2002b,a), a new metric was proposed that falls on a category of its own where the distance between two sequences is based on the angle between the L-tuple count vectors (Eq. 2.15). As these vectors usually have high dimensionality ($K = r^L$, see Eq. 2.1), single value decomposition (SVD) is applied before calculating the angle cosine. Only the dimensions with the higher eigenvalues are used, thus substantially reducing dimensionality with

the additional advantage of filtering some noise from this information. Dimensionality reduction along similar lines has been reported by other authors as being very useful for information retrieval from databases (Berry et al., 1999).

$$\begin{aligned} d_L^{cos}(X,Y) &= \theta_{XY}, \quad \text{where} \\ \cos(\theta_{XY}) &= \frac{(c_L^X)^T \cdot c_L^Y}{\|c_L^X\| \cdot \|c_L^Y\|} = \frac{\sum_{i=1}^K c_{L,i}^X \cdot c_{L,i}^Y}{\sqrt{\sum_{i=1}^K (c_{L,i}^X)^2} \cdot \sqrt{\sum_{j=1}^K (c_{L,j}^Y)^2}} \end{aligned} \tag{2.15}$$

Interestingly, this metric is not sensitive to repetitions, instead returning the difference between the motifs. For example, if a sequence X is compared with its double repetition XX, the vectors c of the counts will have different norms but will have the same direction in space, because $c^X = 2c^{XX}$, causing the angle distance between them to be zero. This property is of fundamental value because it automatically filters repetitions, therefore distinguishing sequences by the different balance of tuple composition only. It is also interesting to note that the distance proposed has strong similarities to the correlation distance d_L^{LCC} (Eq. 2.10). The pairwise cosine values were proposed in the same reports to convert to evolutionary distance, determined from L-tuple counts, as detailed in Eq. 2.16 (Stuart et al., 2002b,a).

$$d_L^{EVOL}(X,Y) = -\ln[(1+\cos\theta_{XY})/2] \tag{2.16}$$

The cross-tabulation of evolutionary distances was then inputted to the NEIGHBOR program (Saitou and Nei, 1987), part of the PHYLIP package (Felsenstein, 1993), used to construct the corresponding phylogenetic trees. The choice of the appropriate L-resolution is further discussed by the proponent whose results suggest it may be specific to the degree of evolutionary divergence. In particular, d_L^{EVOL} was applied to the study of whole mitochondrial genome, and the resulting evolutionary distances were observed to be in agreement with the values previously obtained by other methods. That work put particular emphasis on the dimensionality reduction using the SVD algorithm, which allows a different and interesting interpretation of this metric: by reducing the basis vectors of the representation, the authors are somehow neglecting the main L-tuple composition used, looking for some feature space that conveys a special non-literal representation, in some sense. This can provide a pattern analysis beyond word composition. In principle, the technique could be equally relevant and applied to the preceding metrics.

2.3.2 Resolution free methods

The metrics reviewed above are dependent on a specific resolution or word length of the L-tuples. This problem was solved in some reports cited above by choosing the best discriminant resolution or combining results obtained with arbitrary word-length intervals. Instead, this section reviews alignment-free sequence comparison methods that do not resolve to fixed word-length distance

measures, which represents absolute independence from the assumption of conservation of contiguity. This goal has been pursued following two alternative paths. The first one uses sequence compression as a tool to measure sequence complexity. The extent to which joint compression is more effective than independent compression is used as a measure of similarity. The second approach focuses on the representation of the sequence itself, using iterative functions as bijective maps to continuous, scale-independent formats, where resolution-free comparisons can be pursued.

Universal Sequence Maps (USM)

The pursuit of distance measures independent from L-tuple resolution has been proposed by seeking sequence representations that would themselves be scale independent. Chaos theory, namely as regards the use of iterative functions, is at the foundations of this pursuit. The proposition of iterative functions for the representation of biological sequences is now over a decade old. The original report identified an iterative function for DNA representation, which was named chaos game representation, CGR (Jeffrey, 1990). The recognition that CGR defines a resolution free transition matrix that can be used to derive distance metrics is much more recent (Almeida et al., 2001). That work was later extended and generalized for any order alphabets, thus enabling the study of any discrete sequence, and the new iterative function was renamed Universal Sequence Maps, USM (Almeida and Vinga, 2002). The interesting novel property of the USM bijective mapping is the possibility of accurately represent and summarize any sequence in a continuous multidimensional space at arbitrary resolution (that can be later used to recover sequence context). The comparison of any two unit positions will yield the level of identity between the respective regions in the sequence. For example, the representation of two symbols $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ in USM coordinates can be used to estimate the difference between those symbols in the original sequence (Eq. 2.17).

$$d^{USM}(a, b) = -\log_2(\max_i |a_i - b_i|) \tag{2.17}$$

The USM method can be applied to DNA, proteins and natural language texts but it still is in an experimental development and has not been yet completely tested in challenging sequence sets. It would also be desirable to apply this methodology to multiple comparison and database queries. It should also be noted that the metric proposed, although taking into account symbol context, does not define an overall sequence dissimilarity, like previously reviewed distances.

Kolmogorov Complexity

The use of savings in joint compression as a measure of similarity is founded on information theory and coding, particularly on Kolmogorov Complexity Theory. Similarly to the methods reviewed in the last two sections, this one is also a very recent proposition (Li et al., 2001). The fundamental concept behind

the distance metric proposed is that of algorithmic complexity. In practice, this pursuit requires the use of compression algorithms that are assumed to be efficient. There are presently no absolute measures of algorithmic complexity, which can only be estimated. (For a review of methods see (V'Yugin, 1999).) In that report (Li et al., 2001), sequence compression is performed using the GenCompress software program (Chen et al., 1999), empirically assessing the Kolmogorov complexity, $K(X)$, of a sequence, X by the length of its compressed representation. The conditional complexity is obtained by compressing the juxtaposition of both sequences. The distance measure derived thereof, d^{KC}, detailed in Eq. 2.18, uses the relative decrease in complexity or conditional complexity $K(X|Y)$ as a measure of sequence similarity (Li and Vitanyi, 1997).

$$d^{KC}(X,Y) = 1 - \frac{K(X) - K(X|Y)}{K(XY)} \tag{2.18}$$

The authors demonstrate that d^{KC} satisfies the axioms of a distance function (Eq. 2.5). This method was only tested with mammalian complete mitochondrial genomes (mtDNA), and the distances obtained were observed to be consistent with the known phylogenetic relationships. Despite this method was not yet fully explored, only in a rather limited set of sequences, and the need to estimate the quantities evolved, namely $K(\bullet)$, by a compressing algorithm, it is conceptually attractive and elegant which suggests its further study and extension to higher order alphabets, for example, in comparing proteins.

Recent exploits

The increase in diversity of the newer alignment free distance measures being proposed beyond the framework reviewed here is very apparent as this review is finalized. For example, alignment-independent classification of G-protein coupled receptors (GPCR) based in extracting physical properties of amino acids has been very recently suggested (Lapinsh et al., 2002). This correlation data was processed with multivariate statistical methods, namely Principal Component Analysis (PCA), Partial Least Squares (PLS), autocross-covariance transformations (ACC's), z-scores, in order to weight the individual properties as to correctly classify the proteins studied in super-families. Previous attempts to GPCR classification without alignment were based on the extraction of statistics of communality and specificity for each L-tuple (Daeyaert et al., 1998). These characteristics measure the relative frequency of specific words with regard to the respective super-families.

2.4 Algorithm implementation –NASC-Toolbox

Most of the distance metrics reviewed in this report were coded anew and tested. For that purpose a software toolbox – named Non-aligned sequence comparison (NASC) – was written in MATLAB language and is made publicly available by the authors at http://bioinformatics.musc.edu/NASC. Submission of new distance metrics or more efficient implementation of existing ones to that web-based

File name	Brief description
MANUAL.doc	Toolbox Manual with examples.
EU.txt	Natural language example. EU 10 languages sequences
HUMHBB.txt	Protein example. Translations of the human beta globin region on chromosome 11. [NCBI gi:455025]
thrABC.txt	DNA example. E. coli K12 threonine operon.
cgcria.m cgcgr0.m cgcgr1.m cgcode.m cgle.m cgtp.m	Reads text file, transforms symbol to number, creates USM coordinates.
freqseq.m	Calculates counts and frequencies of L-tuples (or L-words) in sequences previously extracted from file.
overlap.m	Calculates overlap capability of words present.
word_var.m	Variances of L-tuple counts.
word_cov.m	Covariances of L-tuple counts.
distance.m	Calculates different metrics on sequences
nasc.m	Calls all previous functions.
plotdistance.m	Plots all types of distances between chosen sequences.
classif.m	Final sequence classification and dendrogram construction (cluster analysis).
crossd.m	USM cross distances calculation; see also bUSM toolbox (Schwacke and Almeida, 2002).
ang.m	Auxiliary function. Angle between vectors (Euclidean).
h_rel.m	Auxiliary function. Relative entropy between vectors.
isquareform.m	Auxiliary function. Matrix operations.

Table 2.1: Non-aligned sequence comparison (NASC) toolbox MATLAB files

repository is encouraged. The toolbox includes a small manual that explains the algorithms and the use of the functions. It also includes a set of test sequences using different alphabets to exemplify the application of these techniques to sequence classification. Three data sets are included: DNA sequences, protein sequences and natural language – the same text in ten western European idioms with clearly recognizable philology. The following sections are not intended to present an exhaustive study but simply exemplify the potential of these methods on the clustering of sequences. The quantitative analysis of word compositions dissimilarity measures will be fully investigated in Chapter 4.

2.4.1 Matlab functions

Table 2.1 briefly describes the NASC-Toolbox Matlab functions coded and that will be used in the following sections to classify three different sequence types.

2.4.2 DNA

The first example corresponds to the classification of three DNA sequences from the E. coli K12 threonine operon, namely thrA, thrB, thrC. Table 2.2 contains the sequences in FASTA format.

Almost all the metrics tested with several resolution (L) options grouped thrA with thrC in the first step. As an example, the dendrogram obtained with $L = 1$ and Mahalanobis distance is shown in Fig. 2.1. This agrees with other studies, e.g. Almeida et al. (2001).

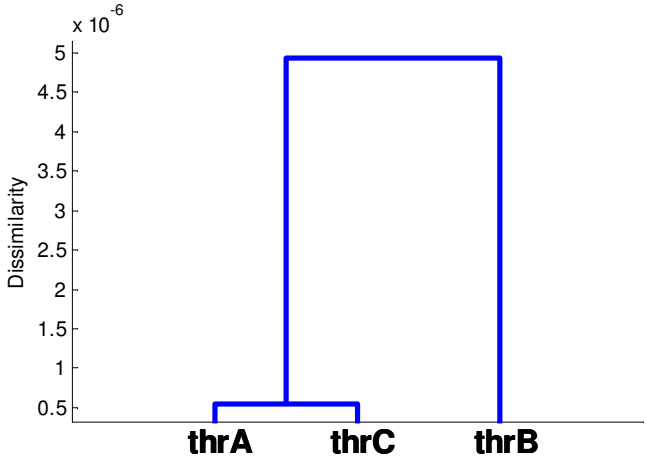


Figure 2.1: Dendrogram with classification of E. coli threonine operon, sequences thrA, thrB and thrC. Mahalanobis dissimilarity (Eq. 2.11) with $L = 1$.

2.4.3 Proteins

The second example tested consists on the classification of five protein sequences from the beta chains of human hemoglobin.

Hemoglobin (Hb) is a protein that is carried by red cells which pick up oxygen in the lungs and delivers it to all peripheral tissues. The hemoglobin molecule is made up of four polypeptide chains (tetramer): two alpha chains (HBA) and two beta chains (HBB).

The alpha chains genes are located in chromosome 16 (not shown). The beta genes (human hemoglobin beta genes or HUMHBB) are located in the chromosome 11 (Locus 11p15.5) as shown in Fig. 2.2. The beta chains change over lifetime, from embryonic to fetal and adult Hb (see figure's legend).

The study of HUMHBB is of major importance in medical science since alterations of beta globin proteins cause several blood diseases such as sickle cell anaemia[2], where only one (the 6th) aminoacid of 146 has been altered due to a mutation on one DNA base (from GAG – Glu to GTG – Val), and thalassemia[2] , in which there is a premature termination of the protein sequence (also due to one substitution from AAG – Lys to TAG – stop codon). Both changes lead to modifications in the folding pattern of the protein and consequently to severe

[2]In portuguese: anemia falciforme and talassémia.

>>E.Coli K12 threonine operon
>thrA
ATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCA
ATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTGAAAA
AACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTTGACGGGACTCGCC
GCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAAAACATGTCCTGC
ATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGC
CATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGG
CATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGG
TGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGC
TGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCG
CGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAG
TTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCTCAAGC
ACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATGGCA
ATGTTCAGCGTTTCTGGTCCGGGGATGAAAGGGATGGTCGGCATGGCGGCGCGCGTCTTTGCAGCGATGTCACGCGCCC
GTATTTCCGTGGTGCTGATTACGCAATCATCTTCCGAATACAGCATCAGTTTCTGCGTTCCACAAAGCGACTGTGTGCG
AGCTGAACGGGCAATGCAGGAAGAGTTCTACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACGG
CTGGCCATTATCTCGGTGGTAGGTGATGGTATGCGCACCTTGCGTGGGATCTCGGCGAAATTCTTTGCCGCACTGGCCC
GCGCCAATATCAACATTGTCGCCATTGCTCAGGGATCTTCTGAACGCTCAATCTCTGTCGTGGTAAATAACGATGATGC
GACCACTGGCGTGCGCGTTACTCATCAGATGCTGTTCAATACCGATCAGGTTATCGAAGTGTTTGTGATTGGCGTCGGT
GGCGTTGGCGGTGCGCTGCTGGAGCAACTGAAGCGTCAGCAAAGCTGGCTGAAGAATAAACATATCGACTTACGTGTCT
GCGGTGTTGCCAACTCGAAGGCTCTGCTCACCAATGTACATGGCCTTAATCTGGAAAACTGGCAGGAAGAACTGGCGCA
AGCCAAAGAGCCGTTTAATCTCGGGCGCTTAATTCGCCTCGTGAAAGAATATCATCTGCTGAACCCGGTCATTGTTGAC
TGCACTTCCAGCCAGGCAGTGGCGGATCAATATGCCGACTTCCTGCGCGAAGGTTTCCACGTTGTCACGCCGAACAAAA
AGGCCAACACCTCGTCGATGGATTACTACCATCAGTTGCGTTATGCGGCGGAAAAATCGCGGCGTAAATTCCTCTATGA
CACCAACGTTGGGGCTGGATTACCGGTTATTGAGAACCTGCAAAATCTGCTCAATGCAGGTGATGAATTGATGAAGTTC
TCCGGCATTCTTTCTGGTTCGCTTTCTTATATCTTCGGCAAGTTAGACGAAGGCATGAGTTTCTCCGAGGCGACCACGC
TGGCGCGGGAAATGGGTTATACCGAACCGGACCCGCGAGATGATCTTTCTGGTATGGATGTGGCGCGTAAACTATTGAT
TCTCGCTCGTGAAACGGGACGTGAACTGGAGCTGGCGGATATTGAAATTGAACCTGTGCTGCCCGCAGAGTTTAACGCC
GAGGGTGATGTTGCCGCTTTTATGGCGAATCTGTCACAACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAGGCCCGTG
ATGAAGGAAAAGTTTTGCGCTATGTTGGCAATATTGATGAAGATGGCGTCTGCCGCGTGAAGATTGCCGAAGTGGATGG
TAATGATCCGCTGTTCAAAGTGAAAAATGGCGAAAACGCCCTGGCCTTCTATAGCCACTATTATCAGCCGCTGCCGTTG
GTACTGCGCGGATATGGTGCGGGCAATGACGTTACAGCTGCCGGTGTCTTTGCTGATCTGCTACGTACCCTCTCATGGA
AGTTAGGAGTCTGA
>thrB
ATGGTTAAAGTTTATGCCCCGGCTTCCAGTGCCAATATGAGCGTCGGGTTTGATGTGCTCGGGGCGGCGGTGACACCTG
TTGATGGTGCATTGCTCGGAGATGTAGTCACGGTTGAGGCGGCAGAGACATTCAGTCTCAACAACCTCGGACGCTTTGC
CGATAAGCTGCCGTCAGAACCACGGGAAAATATCGTTTATCAGTGCTGGGAGCGTTTTTGCCAGGAACTGGGTAAGCAA
ATTCCAGTGGCGATGACCCTGGAAAAGAATATGCCGATCGGTTCGGGCTTAGGCTCCAGTGCCTGTTCGGTGGTCGCGG
CGCTGATGGCGATGAATGAACACTGCGGCAAGCCGCTTAATGACACTCGTTTGCTGGCTTTGATGGGCGAGCTGGAAGG
CCGTATCTCCGGCAGCATTCATTACGACAACGTGGCACCGTGTTTTCTCGGTGGTATGCAGTTGATGATCGAAGAAAAC
GACATCATCAGCCAGCAAGTGCCAGGGTTTGATGAGTGGCTGTGGGTGCTGGCGTATCCGGGGATTAAAGTCTCGACGG
CAGAAGCCAGGGCTATTTTACCGGCGCAGTATCGCCGCCAGGATTGCATTGCGCACGGGCGACATCTGGCAGGCTTCAT
TCACGCCTGCTATTCCCGTCAGCCTGAGCTTGCCGCGAAGCTGATGAAAGATGTTATCGCTGAACCCTACCGTGAACGG
TTACTGCCAGGCTTCCGGCAGGCGCGGCAGGCGGTCGCGGAAATCGGCGCGGTAGCGAGCGGTATCTCCGGCTCCGGCC
CGACCTTGTTCGCTCTGTGTGACAAGCCGGAAACCGCCCAGCGCGTTGCCGACTGGTTGGGTAAGAACTACCTGCAAAA
TCAGGAAGGTTTTGTTCATATTTGCCGGCTGGATACGGCGGGCGCACGAGTACTGGAAAACTAA
>thrC
ATGAAACTCTACAATCTGAAAGATCACAACGAGCAGGTCAGCTTTGCGCAAGCCGTAACCCAGGGGTTGGGCAAAAATC
AGGGGCTGTTTTTTCCGCACGACCTGCCGGAATTCAGCCTGACTGAAATTGATGAGATGCTGAAGCTGGATTTTGTCAC
CCGCAGTGCGAAGATCCTCTCGGCGTTTATTGGTGATGAAATCCCACAGGAAATCCTGGAAGAGCGCGTGCGCGCGGCG
TTTGCCTTCCCGGCTCCGGTCGCCAATGTTGAAAGCGATGTCGGTTGTCTGGAATTGTTCCACGGGCCAACGCTGGCAT
TTAAAGATTTCGGCGGTCGCTTTATGGCACAAATGCTGACCCATATTGCGGGTGATAAGCCAGTGACCATTCTGACCGC
GACCTCCGGTGATACCGGAGCGGCAGTGGCTCATGCTTTCTACGGTTTACCGAATGTGAAAGTGGTTATCCTCTATCCA
CGAGGCAAAATCAGTCCACTGCAAGAAAAACTGTTCTGTACATTGGGCGGCAATATCGAAACTGTTGCCATCGACGGCG
ATTTCGATGCCTGTCAGGCGCTGGTGAAGCAGGCGTTTGATGATGAAGAACTGAAAGTGGCGCTAGGGTTAAACTCGGC
TAACTCGATTAACATCAGCCGTTTGCTGGCGCAGATTTGCTACTACTTTGAAGCTGTTGCGCAGCTGCCGCAGGAGACG
CGCAACCAGCTGGTTGTCTCGGTGCCAAGCGGAAACTTCGGCGATTTGACGGCGGGTCTGCTGGCGAAGTCACTCGGTC
TGCCGGTGAAACGTTTATTGCTGCGACCAACGTGAACGATACCGTGCCACGTTTCCTGCACGACGGTCAGTGGTCACC
CAAAGCGACTCAGGCGACGTTATCCAACGCGATGGACGTGAGTCAGCCGAACAACTGGCCGCGTGTGGAAGAGTTGTTC
CGCCGCAAAATCTGGCAACTGAAAGAGCTGGGTTATGCAGCCGTGGATGATGAAACCACGCAACAGACAATGCGTGAGT
TAAAAGAACTGGGCTACACTTCGGAGCCGCACGCTGCCGTAGCTTATCGTGCGCTGCGTGATCAGTTGAATCCAGGCGA
ATATGGCTTGTTCCTCGGCACCGCGCATCCGGCGAAATTTAAAGAGAGCGTGGAAGCGATTCTCGGTGAAACGTTGGAT
CTGCCAAAAGAGCTGGCAGAACGTGCTGATTTACCCTTGCTTTCACATAATCTGCCCGCCGATTTTGCTGCGTTGCGTA
AATTGATGATGAATCATCAGTAA

Table 2.2: DNA Sequences from E. coli K12 threonine operon

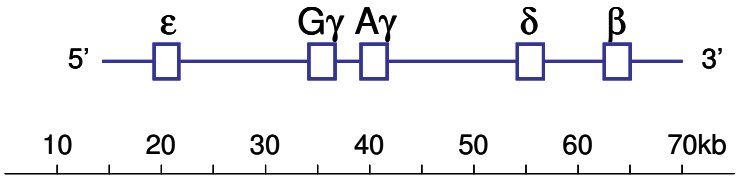


Figure 2.2: Human beta globin genes (HUMHBB) on chromosome 11. The five beta-like globin genes are found within a 45 kb cluster in the following order: ϵ (epsilon) — present in embryonic Hb, later supplanted by fetal and adult Hb $\rightarrow$ G_γ (G-gamma) and A_γ (A-gamma) — expressed in fetal Hb, substituted at birth $\rightarrow$ δ (delta) and β (beta) — expressed in the adult, ca. 3% and 97% respectively of total Hb (together with 2 alfa chains not studied).

```
>>HUMHBB_genes
>HBE1_epsilonglobin
MVHFTAEEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPKVKAHGKKVLTSFGDAIKN
MDNLKPAFAKLSELHCDKLHVDPENFKLLGNVMVIILATHFGKEFTPEVQAAWQKLVSAVAIALAHKYH
>HBG2_Ggammaglobin
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLTSLGDAIKH
LDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTGVASALSSRYH
>HBG1_Agammaglobin
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLTSLGDAIKH
LDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTAVASALSSRYH
>HBD_deltaglobin
MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAH
LDNLKGTFSQLSELHCDKLHVDPENFRLLGNVLVCVLARNFGKEFTPQMQAAYQKVVAGVANALAHKYH
>HBB_betaglobin
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAH
LDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

Table 2.3: Human beta globin sequences in FASTA format.

alteration of the molecule function. This illustrates the importance of tertiary structure in the correct functioning of proteins.

The goal of this section is to classify, by alignment-free methods, the proteins of HUMHBB and compare the results with both the phylogeny of the corresponding genes and with dissimilarity measures based on the multiple alignment of the sequences.

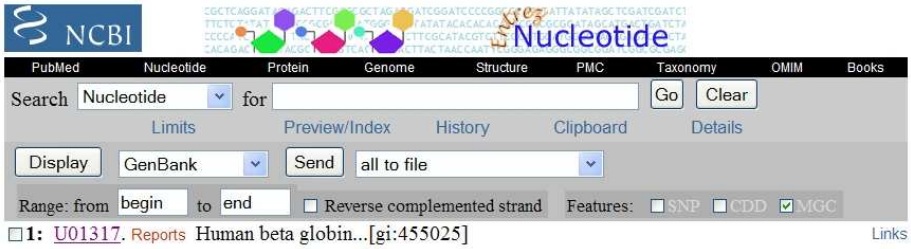
The sequences were extracted from GenBank database. Figure 2.3 shows a snapshot of the corresponding webpage.

After extracting the relevant information, the FASTA format file of the sequences is created (Tab. 2.3).

This FASTA-format file was the input to NASC-Toolbox described above. The options chosen were $L = 1$, which corresponds to using aminoacid frequencies, and cosine dissimilarities , i.e., the angle between vectors as an estimation of dissimilarity between frequency vectors (Eq. 2.15). The dendrogram obtained with this method is represented in Fig. 2.4 and the known phylogeny of the Hb genes is shown in Fig. 2.5.

It is noteworthy the correspondence between both classifications. The clustering obtained with aminoacid frequencies is in accordance with the known phylogenetic relationships, which provides a insightful example of alignment-free techniques for sequence classification.

As an additional comparison procedure the multiple alignment of the sequences was also performed, using program ClustalW (Thompson et al., 1994).




```
LOCUS       HUMHBB                73308 bp    DNA     linear   PRI 13-DEC-2004
DEFINITION  Human beta globin region on chromosome 11.
ACCESSION   U01317 J00179 J00093 J00094 J00096 J00158 J00159 J00160 J00161
            J00162 J00163 J00164 J00165 J00166 J00167 J00168 J00169 J00170
            J00171 J00172 J00173 J00174 J00175 J00177 J00178 K01239 K01890
            K02544 M18047 M19067 M24868 M24886
VERSION     U01317.1  GI:455025
KEYWORDS    Alu repeat; HPFH; KpnI repetitive sequence; RNA polymerase III;
            allelic variation; alternate cap site; beta-1 pseudogene;
            beta-globin; delta-globin; epsilon-globin; gamma-globin; gene
            duplication; globin; polymorphism; promoter mutation; pseudogene;
            repetitive sequence; thalassemia.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 62409 to 62631; 63482 to 63610)
  AUTHORS   Marotta,C.A., Forget,B.G., Weissman,S.M., Verma,I.M.,
            McCaffrey,R.P. and Baltimore,D.
  TITLE     Nucleotide sequences of human globin messenger RNA
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 71 (6), 2300-2304 (1974)
  MEDLINE   74275150
   PUBMED   4135409
REFERENCE   2  (bases 63602 to 63646)
  AUTHORS   Forget,B.G., Marotta,C.A., Weissman,S.M. and Cohen-Solal,M.
  TITLE     Nucleotide sequences of the 3'-terminal untranslated region of
            messenger RNA for human beta globin chain
```

Figure 2.3: GenBank database entry at National Center for Biotechnology Information (NCBI) example – Human beta globin region on chromosome 11.

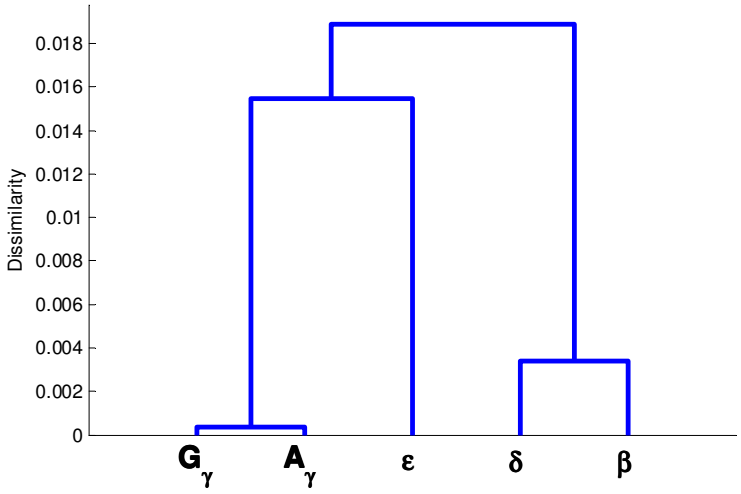


Figure 2.4: Dendrogram of HUMHBB. Legend: 1 – ϵ HBE1_epsilonglobin; 2 – G_γ HBG2_Ggammaglobin; 3 – A_γ HBG1_Agammaglobin; 4 – δ HBD_deltaglobin; 5 – β HBB_betaglobin. The dissimilarity used was the angle between aminoacid frequency vectors ($L = 1$) for each sequence, Eq. 2.15.

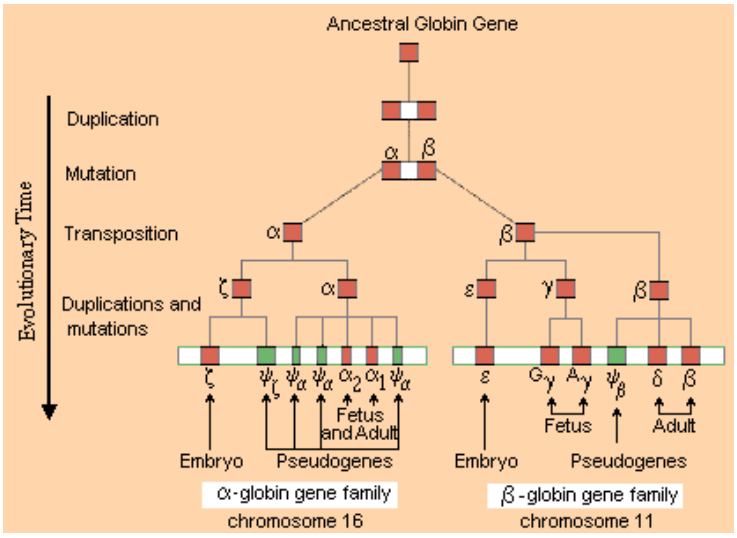


Figure 2.5: Ancestry of the hemoglobin genes. Phylogenetic relationships. In http://www.people.virginia.edu/ rjh9u/globinevolve.html – used with permission © Robert J. Huskey.

Figure 2.6 shows the output of a online application developed at the EBI[3] (Lopez et al., 2003).

```
CLUSTAL W (1.82) multiple sequence alignment


HBG2_Ggammaglobin       MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPK 60
HBG1_Agammaglobin       MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPK 60
HBE1_epsilonglobin      MVHFTAEEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPK 60
HBD_deltaglobin         MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPK 60
HBB_betaglobin          MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
                        * *:* *:*:::.:**.*:**: .***:***************:***:**:..*::****

HBG2_Ggammaglobin       VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG 120
HBG1_Agammaglobin       VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG 120
HBE1_epsilonglobin      VKAHGKKVLTSFGDAIKNMDNLKPAFAKLSELHCDKLHVDPENFKLLGNVMVIILATHFG 120
HBD_deltaglobin         VKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPENFRLLGNVLVCVLARNFG 120
HBB_betaglobin          VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120
                        ********* ::.*.: ::*:** :*: ****************:*****:* :** :**

HBG2_Ggammaglobin       KEFTPEVQASWQKMVTGVASALSSRYH 147
HBG1_Agammaglobin       KEFTPEVQASWQKMVTAVASALSSRYH 147
HBE1_epsilonglobin      KEFTPEVQAAWQKLVSAVAIALAHKYH 147
HBD_deltaglobin         KEFTPQMQAAYQKVVAGVANALAHKYH 147
HBB_betaglobin          KEFTPPVQAAYQKVVAGVANALAHKYH 147
                        ***** :**::**:*:.** **: :**
```

Figure 2.6: Multiple alignment of HUMHBB proteins using ClustalW program - available at http://www.ebi.ac.uk/clustalw/. The similarity is marked in the last row through consensus symbols, which represent the degree of conservation observed in each column: '*' means that the residues or nucleotides in that column are identical in all sequences in the alignment, ':' means that conserved substitutions have been observed, according physicochemical criteria, '.' means that semi-conserved substitutions are observed.

The classification obtained with multiple alignment scores is comparable to the previous results (not shown). The advantage of multiple alignment is to easily recognize the conservation segments and the level of similarity of each aminoacid, which makes this method very insightful in the extraction of substitution and evolutionary processes information.

The previous results show that alignment-free methods were able to cluster the beta globin proteins in accordance with evolutionary information and multiple alignment dissimilarity calculations.

2.4.4 Natural languages

To demonstrate the strength of alignment-free methods for sequences comparison this last example shows an application to natural language texts. This example uses an introductory text to a European Union (EU) site on the web, which is translated in 10 different EU official languages: English (EN), Portuguese (PT), Spanish (SP), Italian (IT), French (FR), German (DE), Dutch (NL), Danish (DA), Finish (FI), Swedish (SW) and Greek (EL), which was excluded from this study given its different alphabet.[4]

The sequences were copied from the web-site http://europa.eu.int, where all the texts are translated in the official EU languages. The main goal of performing the classification of the 10 brands with NASC-Toolbox was to evaluate to what point they were compatible with current evolutionary information from linguistic studies.

[3]The European Bioinformatics Institute (EBI) is a non-profit academic organization that forms part of the European Molecular Biology Laboratory (EMBL).

[4]These were the official languages before the entrance of 10 new countries in 2004.

In the following Tab. 2.4 are represented the 10 sequences used in FASTA format.

The dissimilarity measure applied in this example was the combined Euclidean distance between all pairs of sequences X and Y This is calculated by extracting, for each of the 10 sequences, the 1-tuple to 4-tuple frequencies f_L or counts c_L, which corresponds to the relative abundance of individual letters to 4-mer strings, and calculating all pairwise distances between those vectors (see Eq. 2.8). The composed distance, similar to Eq. 2.13, is given by the following Eq. 2.19:

$$d^{E*} = \sum_{L=1}^{4} d_L^E(X, Y) \tag{2.19}$$

The following Fig. 2.7 represents the dendrogram of the classification using the combined distance defined above.

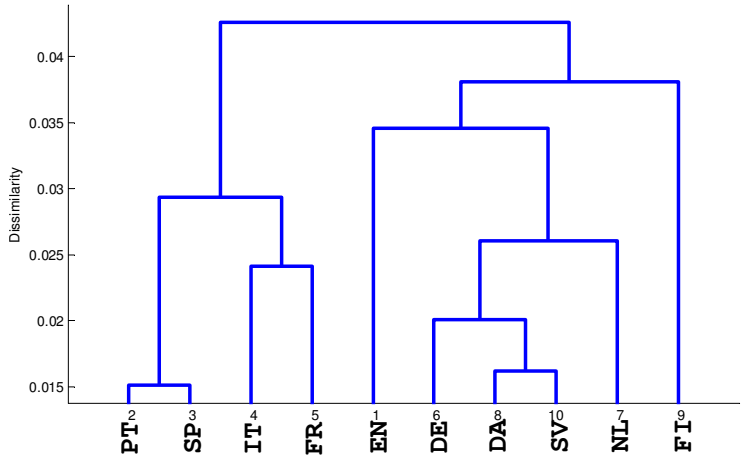


Figure 2.7: Dendrogram with EU languages classification. Classification of EU languages using combined Euclidean distances between all sequences ($L = 1, \ldots, 4$). Legend: PT - Portuguese, SP - Spanish, IT - Italian, FR - French, EN - English, DE - German, DA - Danish, SV - Swedish, NL - Dutch, FI - Finish.

As seen from the figure, there are two main branches corresponding to distinct groups: the Latin branch (also called the Italic or Romance languages) that includes PT, SP, IT and FR, and the Germanic branch, including all the other languages. This classification was even more accurate when joining PT and SP first and then IT and FR. On the other group DA and SV are clustered together first (what corresponds to a known classification of Swedish and Danish, both Scandinavian languages) and with DE after. It is noteworthy that FI is far apart, which is also known in Linguistic studies since Finish belongs to the Uralic family while all the other languages studied belong to the Indo-European family.[5]

[5]Feature not totally captured with this classification since FI was clustered with the Germanic branch instead of being in a separate group.

>>EU
>EnglishEN
The History of the European Union, presents the chronology of important accomplishments of the EU and its institutions. It is a selection of events, updated on a monthly basis. This selection is based on the Bulletin of the European Union (published 10 times a year), and is fully revised once a year with the annual General Report on the Activities of the European Union. From Robert Schuman's declaration of 1950 to the first enlargement waves in the 70's and the 80's, from the establishment of the Single Market in 1993 to the introduction of the euro notes and coins on January 1st, 2002, and the opening of enlargement negotiations with the countries of Eastern and Central Europe.
>PortuguesePT
A História da União Europeia apresenta a cronologia das mais importantes realizações da UE e das suas instituições. Trata-se de uma selecção de acontecimentos, actualizada numa base mensal. Esta selecção baseia-se no Boletim da União Europeia (publicado 10 vezes por ano) e revista uma vez por ano em função do Relatório Geral das Actividades da União Europeia. Da Declaração de Robert Schuman de 1950 até à primeira vaga de alargamentos nas décadas de 70 e de 80 e da criação do Mercado Interno em 1993 à introdução das notas e moedas em euros em 1 de Janeiro de 2002 e ao lançamento das negociações de alargamento com os países da Europa Central e Oriental.
>SpanishSP
La Historia de la Unión Europea presenta la cronología de los acontecimientos más importantes protagonizados por la UE y sus instituciones. Es una selección de acontecimientos, actualizada mensualmente, basada en el Boletín de la Unión Europea(que se publica 10 veces al año) y revisada completamente una vez al año de acuerdo con el Informe General sobre las Actividades de la Unión Europea. En ella puede encontrarse desde la declaración de Robert Schuman de 1950 hasta las primeras oleadas de ampliación en las décadas de los 70 y los 80, y desde el establecimiento del Mercado Único en 1993 hasta la introducción de los billetes y monedas de euro el 1 de enero de 2002 y el inicio de las negociaciones de ampliación con los países de Europa Central y Oriental.
>ItalianIT
La Storia dell'Unione europea è una presentazione cronologica dei principali eventi che hanno segnato la vita dell'Unione europea e delle sue istituzioni. È una selezione di evento che viene aggiornata mensilmente sulla base del Bollettino dell'Unione europea (10 numeri all'anno) e rivista ogni anno alla luce della Relazione generale sull'attività dell'Unione europea, e che va dalla dichiarazione di Robert Schuman del 1950 ai primi allargamenti degli anni Settanta e Ottanta, passando per la instaurazione del mercato unico nel 1993, l'immissione in circolazione delle monete e banconote in euro il 1 gennaio 2002 e il varo dei negoziati di adesione con i paesi dell'Europa centrale e orientale.
>FrenchFR
L'histoire de l'Union européenne donne un aperçu chronologique des principales réalisations de l'UE et de ses institutions. Il s'agit d'une sélection d'événements mise à jour tous les mois. Cette sélection repose sur le Bulletin de l'Union européenne (publié dix fois par an) et est entièrement revue chaque année sur la base du Rapport général sur l'activité de l'Union européenne. Les événements repris vont de la déclaration de Robert Schuman, en 1950, aux premiers élargissements des années 70 et 80, de la mise en place du marché unique, en 1993, à l'introduction des billets et des pièces en euros, le 1er janvier 2002, et à l'ouverture des négociations d'adhésion avec les pays d'Europe centrale et orientale.

Table 2.4: EU languages

>GermanDE Die Geschichte der Europäischen Union enthält eine chronologische Übersicht über die wichtigsten Errungenschaften der EU und ihrer Organe. Es handelt sich um eine monatlich aktualisierte Auswahl von Ereignissen, die sich auf das Bulletin der Europäischen Union (erscheint zehnmal im Jahr) stützt und im Rahmen des Gesamtberichts über die Tätigkeit der Europäischen Union einmal im Jahr vollständig aktualisiert wird. Sie reicht von der Erklärung Robert Schumans aus dem Jahr 1950 bis zu den ersten Erweiterungen in den 70er und 80er Jahren, von der Errichtung des Binnenmarktes im Jahr 1993 bis zur Einführung der Euro-Banknoten und -Münzen am 1. Januar 2002 und der Aufnahme der Erweiterungsverhandlungen mit den Ländern Mittel- und Osteuropas. >NetherlandsNL De geschiedenis van de Europese Unie bevat een chronologisch overzicht van de belangrijkste stappen in de ontwikkeling van de EU en haar instellingen. Het is een selectie van gebeurtenissen die maandelijks wordt bijgewerkt. Deze selectie is gebaseerd op het Bulletin van de Europese Unie, dat 10 keer per jaar verschijnt, en op het Algemeen verslag over de werkzaamheden van de Europese Unie, waarin eenmaal per jaar al deze informatie wordt samengevat. Het overzicht loopt van de verklaring van Robert Schuman van 1950, via de eerste uitbreidingsgolven in de jaren zeventig en tachtig en de totstandbrenging van de interne markt in 1993, tot de invoering van de euromunten en - biljetten op 1 januari 2002 en het begin van de toetredingsonderhandelingen met de landen van Midden- en Oost-Europa. >DanishDA I Den Europæiske Unions historie finder man en kronologisk oversigt over de vigtigste resultater, EU og EU's institutioner har opnået. Det er et udvalg af begivenheder, som ajourføres hver måned. Udvælgelsen sker på grundlag af Bulletinen for Den Europæiske Union (udkommer 10 gange om året) og revideres fuldstændigt en gang om året i den almindelige beretning om Den Europæiske Unions virksomhed. Fra Robert Schumans erklæring i 1950 til de første bølger af udvidelser i 1970'erne og 1980'erne, fra oprettelsen af det indre marked i 1993 til indførelsen af eurosedler og -mønter den 1. januar 2002 og indledningen af udvidelsesforhandlinger med landene i Øst- og Centraleuropa. >FinishFI Tässä Euroopan unionin historiassa käydään läpi EU:n ja sen toimielinten tärkeimmät saavutukset aikajärjestyksessä. Kyseessä on kuukausittain päivitettävä kooste tapahtumista. Kooste perustuu Euroopan unionin tiedotteeseen (ilmestyy 10 kertaa vuodessa). Kerran vuodessa tehtävän perusteellisemman tarkistuksen pohjana on vuosittain ilmestyvä Yleiskertomus Euroopan unionin toiminnasta. Katsaus ulottuu Ranskan ulkoministerin Robert Schumanin vuonna 1950 antamasta, Euroopan hiili- ja teräsyhteisön perustamiseen johtaneesta julistuksesta ensimmäisiin laajentumiskierroksiin 1970- ja 1980-luvulla, yhtenäismarkkinoiden perustamiseen vuonna 1993, euroseteleiden ja -kolikoiden käyttöönottoon 1. tammikuuta 2002 sekä laajentumisneuvotteluiden käynnistämiseen Keski- ja Itä-Euroopan maiden kanssa. >SwedishSV Europeiska unionens historia ger en kronologisk översikt över viktiga händelser i EU:s och EU-institutionernas historia. Det är ett urval som uppdateras varje månad. Uppgifterna är hämtade från Europeiska unionens bulletin, som utkommer tio gånger om året, och från den årliga Allmänna rapporten om Europeiska unionens verksamhet. I översikten hittar du allt, från Robert Schumans deklaration 1950, de första utvidgningsomgångarna på sjuttio- och åttiotalen och upprättandet av den inre marknaden 1993 till övergången till mynt och sedlar i euro den 1 januari 2002 och de pågående anslutningsförhandlingarna med länderna i Central- och Österuropa.

Table 2.4: EU languages (cont.)

2.5 Conclusions

Sequence comparison by alignment has both fundamental and computational limitations. The conservation of contiguity underlying alignment is at odds with genetic recombination, which includes shuffling subgenomic DNA fragments. This limitation is particularly clear by recalling that, regardless of the progress in the identification of scoring matrices, alignment fails to recognize proteomic sequences with less than 20% sequence identity. In addition, optimal alignment is computationally too heavy for efficient querying the sharply inflating genomic and proteomic public databases. The increasing awareness of those limitations is driving the proposition of a diversity of new foundations for alignment-free sequence analysis, hereby reviewed. The diversity of theoretical foundations explored by the reports reviewed here ranges from linear algebra and statistics, to information theory, Kolmogorov complexity and chaos theory. The recent abundance of successful applications of alignment-free sequence analysis, and the increasing focus on practical implementations makes it a safe prediction that the next few years will see some of them become widely used for functional annotation and phylogenetic study.

2.6 Acknowledgements

The authors thankfully acknowledge the financial support by grants SFRH/-BD/3134/2000 and Sapiens/34794/99 of Fundação para a Ciência e Tecnologia of the Portuguese Ministry of Science and Technology (FCT/MCT).

2.7 References

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., and Fletcher, M. (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics*, 17(5):429–437. 46, 49

Almeida, J. S. and Vinga, S. (2002). Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*, 3(1):6. 46

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215:403–410. 35, 41

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402. 35

Ash, R. B. (1990). *Information Theory*. Dover Publications, New York. 38

Attwood, T. K. (2000). Genomics: The babel of bioinformatics. *Science*, 290(5491):471–3. 34

Berry, M. W., Drmac, Z., and Jessup, E. R. (1999). Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2):335–362. 45

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci USA*, 83(14):5155–9. 38, 39

Blaisdell, B. E. (1989a). Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. *J Mol Evol*, 29(6):538–47. 40

Blaisdell, B. E. (1989b). Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. *J Mol Evol*, 29(6):526–37. 40

Burke, J., Davison, D., and Hide, W. (1999). d2-cluster: a validated method for clustering est and full-length cdna sequences. *Genome Res*, 9(11):1135–42. 41

Burke, J., Wang, H., Hide, W., and Davison, D. B. (1998). Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res*, 8(3):276–90. 41

Carpenter, J. E., Christoffels, A., Weinbach, Y., and Hide, W. A. (2002). Assessment of the parallelization approach of d2-cluster for high-performance sequence clustering. *J Comput Chem*, 23(7):755–7. 41

Chen, X., Kwong, S., and Li, M. (1999). A compression algorithm for DNA sequences and its applications in genome comparison. *Genome Inform Ser Workshop Genome Inform*, 10:51–61. 47

Christoffels, A., van Gelder, A., Greyling, G., Miller, R., Hide, T., and Hide, W. (2001). STACK: Sequence tag alignment and consensus knowledgebase. *Nucleic Acids Res*, 29(1):234–8. 41

Coghlan, A., Mac Donaill, D. A., and Buttimore, N. H. (2001). Representation of amino acids as five-bit or three-bit patterns for filtering protein databases. *Bioinformatics*, 17(8):676–85. 41

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York. 38

Daeyaert, F., Moereels, H., and Lewi, P. J. (1998). Classification and identification of proteins by means of common and specific amino acid n-tuples in unaligned sequences. *Comput Methods Programs Biomed*, 56(3):221–33. 47

Davison, D. and Burke, J. (2001). Brute force estimation of the number of human genes using est clustering as a measure. *IBM J. Res. & Dev.*, 45(3/4):439–447. 41

Dayhoff, M. O., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In Dayhoff, M. O., editor, *Atlas of protein sequence and structure*, volume 5 - supplement 3, pages 345–352. National Biomedical Research Foundation, Washington, D.C. 35

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press. 34

Everitt, B., Landau, S., and Leese, M. (2001). *Cluster Analysis*. Arnold, London, 4th edition. 35

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package)*. Department of Genetics, University of Washington, Seattle, version 3.5c – Distributed by the author. 40, 45

Fichant, G. and Gautier, C. (1987). Statistical method for predicting protein coding regions in nucleic acid sequences. *Comput Appl Biosci*, 3(4):287–95. 41

Fuchs, R. (2002). From sequence to biology: the impact on bioinformatics. *Bioinformatics*, 18(4):505–6. 34

Gentleman, J. F. and Mullin, R. C. (1989). The distribution of the frequency of occurrence of nucleotide subsequences, based on their overlap capability. *Biometrics*, 45(1):35–52. 37, 43

Gibbs, A. J., Dale, M., Kinns, H. R., and MacKenzie, H. G. (1971). The transition matrix method for comparing sequences; its use in describing and classifying proteins by their amino acis sequence. *Systematic Zoology*, 20:417–425. 41

Giladi, E., G.Walker, M., Wang, J. Z., and Volmuth, W. (2002). SST: an algorithm for finding near-exact sequence matches in time proportional to the logarithm of the database size. *Bioinformatics*, 18(6):873–877. 41

Gisiger, T. (2001). Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biol Rev Camb Philos Soc*, 76(2):161–209. 40

Gordon, A. (1999). *Classification*. Chapman & Hall/CRC, Boca Raton, 2nd edition. 35

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J Mol Biol*, 162:705–708. 34

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press. 34

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, 89(22):10915–9. 35

Hide, W., Burke, J., Christoffels, A., and Miller, R. (1997). A novel approach towards a comprehensive consensus representation of the expressed human genome. In Miyano, S. and Takagi, T., editors, *Genome Informatics*, pages 187–196. Universal Academy Press, Tokyo, Japan. 41

Hide, W., Burke, J., and Davison, D. B. (1994). Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J Comput Biol*, 1(3):199–215. 41

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res*, 18(8):2163–2170. 46

Kullback, S. (1968). *Information theory and statistics*. Dover Publications, New York. 37

Lapinsh, M., Gutcaits, A., Prusis, P., Post, C., Lundstedt, T., and Wikberg, J. E. (2002). Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci*, 11(4):795–805. 47

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P., and Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–54. 46, 47

Li, M. and Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 2nd edition. 47

Lopez, R., Robinson, S., Kibria, A., Harte, N., Patel, G., Harper, R., Quevillon, E., Silventoinen, V., Kallio, K., and Jokinen, P. (2003). The European Bioinformatics Institute web site: a new view. *Bioinformatics*, 19(4):546–547. 54

Lynch, M. (2002). Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A*, 99(9):6118–23. 34

Miller, R. T., Christoffels, A. G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T. R., and Hide, W. A. (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res*, 9(11):1143–55. 41

Mironov, A. A. and Alexandrov, N. N. (1988). Statistical method for rapid homology search. *Nucleic Acids Res*, 16(11):5169–73. 42

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48:443–453. 34

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, 183:63–98. 35

Pearson, W. R. (2000). Protein sequence comparison and protein evolution. *Tutorial – ISMB2000*. 34

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 85(8):2444–8. 35, 41

Petrilli, P. (1993). Classification of protein sequences by their dipeptide composition. *Comput Appl Biosci*, 9(2):205–9. 42

Petrilli, P. and Tonukari, N. J. (1997). PFDB: a protein families database for macintosh computers. the effectiveness of its organization in searching for protein similarity. *J Protein Chem*, 16(7):713–20. 42

Pevzner, P. A. (1992). Statistical distance between texts and filtration methods in sequence comparison. *Comput Appl Biosci*, 8(2):121–7. 40

Régnier, M. (1998). A unified approach to word statistics. In Press, A., editor, *Proceedings of the second annual international conference on Computational molecular biology*, pages 207–213, New York, United States. 37

Reichhardt, T. (1999). It's sink or swim as a tidal wave of data approaches. *Nature*, 399(6736):517–20. 34

Reinert, G., Schbath, S., and Waterman, M. S. (2000). Probabilistic and statistical properties of words: an overview. *J Comput Biol*, 7(1–2):1–46. 37, 43

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–25. 45

Schott, J. R. (1997). *Matrix analysis for statistics*. John Wiley, New York. 37

Schwacke, J. and Almeida, J. S. (2002). Efficient boolean implementation of universal sequence maps (bUSM). *BMC Bioinformatics*, 3(1):28. 48

Searls, D. B. (2001). Reading the book of life. *Bioinformatics*, 17(7):579–80. 34

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656. 37

Siegmund, D. and Yakir, B. (2000). Approximate p-values for local sequence alignments. *The Annals of Statistics*, 28(3):657–680. 35

Sitnikova, T. L. and Zharkikh, A. A. (1993). Statistical analysis of L-tuple frequencies in eubacteria and organelles. *Biosystems*, 30(1-3):113–35. 40

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147:195–197. 34

Solovyev, V. V. and Makarova, K. S. (1993). A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. *Comput Appl Biosci*, 9(1):17–24. 42, 44

Storey, J. D. and Siegmund, D. (2001). Approximate p-values for local sequence alignments: numerical studies. *J Comput Biol*, 8(5):549–56. 35

Strang, G. (1988). *Linear Algebra and Its Applications*. International Thomson Publishing, 3rd edition. 37

Stuart, G. W., Moffett, K., and Baker, S. (2002a). Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, 18(1):100–8. 44, 45

Stuart, G. W., Moffett, K., and Leader, J. J. (2002b). A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol*, 19(4):554–62. 44, 45

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80. 51

Torney, D. C., Burks, C., Davison, D., and Sirotkin, K. M. (1990). Computation of d2: a measure of sequence dissimilarity. G.I. Bell and T.G. Marr (eds). In *Computers and DNA: the proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop, held December 12 to 16, 1988 in Santa Fe, New Mexico*, pages 109–125, Redwood City, Calif. Addison-Wesley. 40

URL (2002). http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html. *NCBI*. 40

van Heel, M. (1991). A new family of powerful multivariate statistical sequence analysis techniques. *J Mol Biol*, 220(4):877–87. 41

V'Yugin, V. V. (1999). Algorithmic complexity and stochastic properties of finite binary sequences. *The Computer Journal*, 42(4):294–317. 47

Waterman, M. S. (1995). *Introduction to computational biology: maps, sequences, and genomes*. Chapman & Hall/CRC, Boca Raton, Fla. 34, 37

Wu, T. J., Burke, J. P., and Davison, D. B. (1997). A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, 53(4):1431–9. 43, 44

Wu, T. J., Hsieh, Y. C., and Li, L. A. (2001). Statistical measures of DNA sequence dissimilarity under markov chain models of base composition. *Biometrics*, 57(2):441–8. 44

Zhang, Y. X., Perry, K., Vinci, V. A., Powell, K., Stemmer, W. P., and del Cardayre, S. B. (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature*, 415(6872):644–6. 34

Zharkikh, A. A. and Rzhetsky, A. (1993). Quick assessment of similarity of two sequences by comparison of their *L*-tuple frequencies. *Biosystems*, 30(1-3):93–111. 40

Chapter 3

Universal sequence map (USM) of arbitrary discrete sequences

Published in: Almeida, JS. and Vinga, S. (2002). Universal Sequence Map (USM) of arbitrary discrete sequences. BMC Bioinformatics *3:6.*

Background For over a decade the idea of representing biological sequences in a continuous coordinate space has maintained its appeal but not been fully realized. The basic idea is that any sequence of symbols may define trajectories in the continuous space conserving all its statistical properties. Ideally, such a representation would allow scale independent sequence analysis – without the context of fixed memory length. A simple example would consist on being able to infer the homology between two sequences solely by comparing the coordinates of any two homologous units.

Results We have successfully identified such an iterative function for bijective mapping of discrete sequences into objects of continuous state space that enable scale-independent sequence analysis. The technique, named Universal Sequence Mapping (USM), is applicable to sequences with an arbitrary length and arbitrary number of unique units and generates a representation where map distance estimates sequence similarity. The novel USM procedure is based on earlier work by these and other authors on the properties of Chaos Game Representation (CGR). The latter enables the representation of 4 unit type sequences (like DNA) as an order free Markov chain transition table. The properties of USM are illustrated with test data and can be verified for other data by using the accompanying web-based tool (http://bioinformatics.musc.edu/∼jonas/usm/).

Conclusion USM is shown to enable a statistical mechanics approach to sequence analysis. The scale independent representation frees sequence analysis from the need to assume a memory length in the investigation of syntactic rules.

3.1 Background

For over a decade the idea of representing biological sequences in a continuous coordinate space has maintained its appeal but not been fully realized (Román-Roldán et al., 1996; Nady, 1994; Tino, 1999). The basic idea is that sequences of symbols, such as nucleotides in genomes, aminoacids in proteomes, repeated sequences in Multi-Locus Sequence Typing (MLST) (Enright et al., 2000), words in languages or letters in words, would define trajectories in this continuous space conserving the statistical properties of the original sequences (Tino, 1999; Jeffrey, 1990; Hill and Singh, 1997; Forte et al., 1998; Fiser et al., 1994; Deschavanne et al., 1999). Accordingly, the coordinate position of each unit would uniquely encode for both its identity and its context, i.e. the identity of its neighbors (Roy et al., 1998). Ideally, the position should be scale-independent, such that the extraction of the encompassing sequence can be performed with any resolution, leading to an oligomer of arbitrary length. The pioneer work by Jeffrey published in 1990 (Jeffrey, 1990) achieved this for genomic sequences by using the Chaos Game Representation technique (CGR), defining a unit-square where each corner corresponds to one of the 4 possible nucleotides. Subsequent work further explored the properties of CGR of biological sequences, but two main obstacles prevented the realization of its early promise – lack of scalability with regard to the number of possible unique units and inability to represent succession schemes. Meanwhile, Markov chain theory already offered a solid foundation for the identification of discrete spaces to represent sequences as cross-tabulated conditional probabilities – Markov transition tables. This Bayesian technique is widely explored in bioinformatic applications seeking to measure homology and align sequences (Durbin et al., 1998). In a recent report (Almeida et al., 2001) we have shown that, for genomic sequences, Markov tables are in fact a special case of CGR, contrary to what had been suggested previously (Goldman, 1993). This raised the prospect of an advantageous use of iterative maps as state spaces not only for representation of sequences but also to identify scale independent stochastic models of the succession scheme. That work (Almeida et al., 2001) is hereby extended and further generalized to be applicable to sequences with arbitrary numbers of unique component units, without sacrificing the inverse correlation between distance in the map and sequence similarity independent of position. Accordingly, the technique is named Universal Sequence Map (USM).

3.2 Results

The Results are divided in two sections. The first section presents the foundations for identifying an iterative function with the desired properties. The second section describes algorithm implementation illustrated with a sample data set. Both sections are best understood by using the accompanying web-based tool (see Abstract for address) where the different steps of the procedure can be verified and reproduced with the test data or the reader's own data.

3.2.1 Conceptual foundations

The USM generalization proposed here is achieved by observing two stipulations: A – alternative units in the iterative map are positioned in distinct corners of unit block structures; and B – sequence processing is bi-directional.

Basis for USM generalization:

A. Each unique unit is referenced in the map for positions that are at equal n-distances from each other, and possibly, but not necessarily, defining a complete block structure (Tino, 1999). n-distances are defined as the maximum distance along any dimension, e.g. n-distance between $[a_1, a_2, \ldots, a_n]$ and $[b_1, b_2, \ldots, b_n]$ is $\max(|b_1 - a_1|, |b_2 - a_2|, \ldots, |b_n - a_n|)$, see also Eq. 3.3. It will be shown that this stipulation leads to the definition of spaces where distance is inversely proportional to sequence similarity, independent of position. In this respect, USM departs from previous attempts to generalize chaos game representation that conserve the bi-dimensionality of the original CGR representation (Fiser et al., 1994; Basu et al., 1997; Pleissner et al., 1997; Solovyev et al., 1993).

B. The iterative positioning is performed in both directions. Therefore, there will be two sets of coordinates, the result of forward and backward iterative operations. It will be shown that, by adding backward and forward map distances between two positions, the number of identical units in the encompassing sequences can be extracted directly from the USM coordinates. As a consequence, two arbitrary positions can be compared, and the number of contiguous similar units is extracted by an algebraic operation that relies solely on the USM coordinates of those very two positions.

3.2.2 Implementation of USM algorithm

The algorithm will be first illustrated for the first and last stanzas of Wendy Cope's poem "The Uncertainty of the Poet" (Cope, 1992), respectively, "I am a poet. I am very fond of bananas." and "I am of very fond bananas. Am I a poet?". The procedure includes four steps:

1. Identification of unique sequence units – e.g. these two stanzas have 19 unique characters, (Tab. 3.1), i.e. $uu = 19$.

2. Replacement of each unique unit (in this case units are alphabetic characters) by a unique binary number – e.g. in Tab. 3.1 each of the 19 unique units is replaced by its rank order minus one, represented as a binary number. Other arrangements are possible leading to the same final result as discussed below. The minimum number of dimensions necessary to accommodate uu unique units, n, is the upper integer of the length of its binary representation: $n = \lceil \log_2(uu) \rceil$. For W. Cope's stanzas, $n = \lceil \log_2(19) \rceil = 5$. The binary reference coordinates for the unique units are defined by the numerals of the binary code – for example, a will be assigned to the position $U_{\text{'a'}} = [0, 0, 1, 0, 1]$. Each symbol is represented as a corner in a n-dimensional cube (Tab. 3.1). The purpose of these first two steps is to guarantee that the reference positions for each unique sequence unit component are equidistant (stipulation A) in the n-metric defined above. Any other procedure resulting in equidistant unique positions will lead to the same final results independently of the actual binary numbers used or the number of dimensions used to contain them.

Unit	Bin.Code
	00000
.	00001
?	00010
A	00011
a	00101
b	00110
d	00111
e	01000
f	01001
I	00100
m	01010
n	01011
o	01100
p	01101
r	01110
s	01111
t	10000
v	10001
y	10010

Table 3.1: Binary codes for the 19 possible units occurring in the two stanzas. The first unit is a space character " ".

3. The CGR procedure (Jeffrey, 1990) (Eq. 3.1) is applied independently to each coordinate, $j = 1, 2, \ldots, n$, for each unit, i, in the sequence of length k, $u_j(i)$ with $i = 1, 2, \ldots, k$, and starting with a random map position taken from a uniform distribution in $[0, 1]^n$, i.e. $Unif([0, 1]^n)$. The random seed is not fundamentally different from using the middle position in the map as is conventional in CGR and it has the added feature that it prevents the invalidation of the inverse logarithmic proportionality of n-distance to sequence similarity (Almeida et al., 2001) for sequences that start or end with the same motif.

For a sequence with k units, the USM positions $i = 1, \ldots, k$ for the $j = 1, \ldots, n$ dimensions are determined as follows:

$$\begin{cases} USM_j^{(0)} \sim Unif([0,1]) \\ USM_j^{(i)} = USM_j^{(i-1)} + \frac{1}{2}\left(u_j^{(i)} - USM_j^{(i-1)}\right) = \frac{1}{2}USM_j^{(i-1)} + \frac{1}{2}u_j^{(i)} \end{cases} \tag{3.1}$$

4. The previous step generated k positions in a n-dimension space by processing the sequence forward (Eq. 3.1). This subsequent step adds an additional set of n dimensions by implementing the same procedure backward (Eq. 3.2), again starting at random positions for each coordinate. Consequently the first n dimensions of USM will be referred to as defining a *forward map* and the second set of n dimensions will define a *backward map*. Put together, the bi-directional USM map defines a $2n$-unit block structure.

The n additional backward coordinates are determined as follows:

$$\begin{cases} USM_{n+j}^{(k+1)} \sim Unif([0,1]) \\ USM_{n+j}^{(i)} = \frac{1}{2} USM_{n+j}^{(i+1)} + \frac{1}{2} u_j^{(i)} \end{cases} \tag{3.2}$$

The forward USM map for genomic sequences, where $uu = 4$, and, consequently, $n = 2$, is the same as the result generated by CGR. However, by freeing the iterative map from the dual-dimensional constraint of conventional CGR, the USM forward map alone achieved the goal of producing a scale independent representation of sequences of arbitrary number of unique units. These properties will be briefly illustrated with W. Cope's example. The 16th unit of the first stanza, "I am a poet. I am very fond of bananas.", has USM coordinates $USM_{[1,\ldots,2n]}^{(16)}$ =[0.02, 0.01, 0.63, 0.00, 0.53, 0.07, 0.30, 0.52, 0.27, 0.57]. The first $n = 5$ coordinates, the position in the forward map, can now be used, by reversing Eq. 3.1 (Almeida et al., 2001; Goldman, 1993), not only to extract the identity the unit $i = 16$ but also the identity of the preceding units:

–using forward coordinates alone $(0.0156, 0.0138, 0.6314, 0.0001, 0.5338)$

$\rightarrow$ 00100 00000 00101 01010 00000 00101 00000 01101 01100 01000 10000 00001 00000 00100 00000 00101 $\rightarrow$ *I am a poet. I a*

The same procedure can be applied to the remaining $n = 5$ coordinates, the position in the backward map, to extract the identity of the succeeding units, now ordered backwards.

–using backward coordinates alone $(0.0703, 0.3004, 0.5169, 0.2742, 0.5652)$

$\rightarrow$ 00011 01111 00101 01011 00101 01011 00101 00110 00000 01001 01100 00000 00111 01011 01100 00000 01001 10010 01110 01000 10001 00000 01010 01010 $\rightarrow$ *.sananab fo dnof yrev m*

The length of the sequence that can be recovered from a position in the CGR or USM space is only as long as the resolution, in bits, of the coordinates themselves. In addition, the relevance of these iterative techniques is not associated with the property of recovering sequences as much as with the ability to recover the succession schemes, e.g. the Markov probability tables. It has been recognized for almost a decade that the density of positions in unidirectional, bidimensional, iterated CGR maps (e.g. of genomic sequences, $uu = 4 \rightarrow n = 2$) defines a Markov table (Almeida et al., 2001; Goldman, 1993). The complete accommodation of Markov chains in unidirectional USM (i.e. either forward or backward, which is an equivalent to a multidimensional solution for CGR) can be quickly established by noting that the identity of a quadrant is set by its middle coordinates (Goldman, 1993). In order to extract the Markov format, for an arbitrary integer order *ord*, each of the two n-unit hypercubes, the set of n or backward coordinates, would be divided in $q = 2^{n \cdot (ord+1)}$ equal quadrants and the quadrant frequencies rearranged (Almeida et al., 2001). The use of quadrant to designate what is in fact a sub-unit hypercube is a consonance with the preceding work on bidimensional CGR maps (Almeida et al., 2001), where it was shown that since any number of subdivisions can be considered in a continuous domain, the density distribution becomes an order-free Markov table that accommodates both integer and fractal memory lengths. The extraction of Markov chain transition tables from USM representations, both forward and backward, is included in the accompanying web-based application (see Abstract).

Above, the USM procedure was shown to allow for the representation of sequences as multidimensional objects without loss of identity or context. These objects can now be analyzed to characterize the sequences for quantities such as similarity between segments or entropy (Román-Roldán et al., 1994; Oliver et al., 1993) within the sequence. In Figure 3.1 the 10-dimensional object defined by

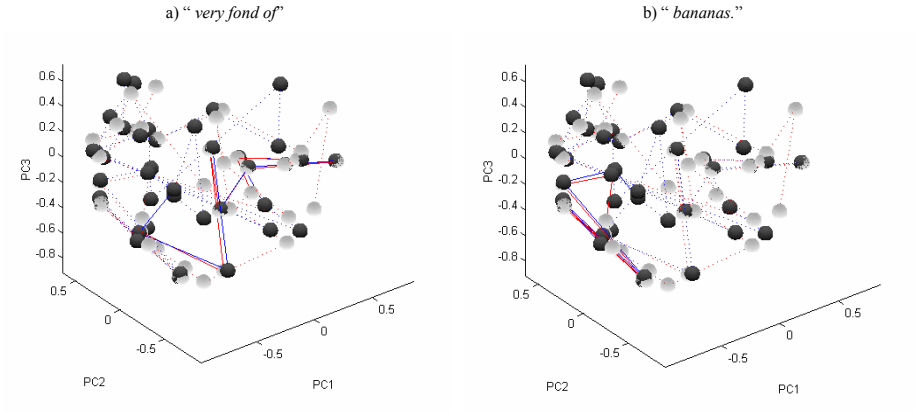


Figure 3.1: Representation of the USM of the two stanzas,respectively dark and light spheres connected by dashed lines, in a reduced 3-dimension space obtained using the first three principal components, $PC_{1,2,3}$. In a) the units corresponding to the segment "very fond of" in both stanzas are connected by solid lines. The procedure is repeated in b) for the segment "bananas.". These figures illustrate the property that similar segments converge in the USM representation, which is reflected by the docking of homologous units. The factorization for dimensionality reduction serves visualization purposes only. The variance represented by each of the three principal components is 40%, 13% and 11%, respectively.

the USM positions of the two stanzas was projected in 3-dimensions by principal component analysis. The dimensionality reduction by principal factor extraction has visualization purposes only. As established above, the minimum necessary dimensionality of the USM state space is set by the binary logarithm of the number of unique units. Nevertheless, the sequence variance associated with each component is provided in the figure legend. In Figure 3.1a, the segments "very fond of" in the two stanzas are linked by solid lines to highlight the fact that sequence similarity is reflected by spatial proximity of USM coordinates. The representation is repeated in Fig. 3.1b with solid lining of the segment "bananas". The matching of the two segments of the second stanza (light) to the similar segments of the first stanza (dark) is, again, visually apparent.

The USM algorithm determines that similar sequences, or segments of sequences, will have converging iterated trajectories: the distance will be cut in half for every consecutive similar unit.[1] This property was noticed before for CGR of genomic sequences (Almeida et al., 2001), and will be further explored here for USM generalization. In that preceding work it was shown that the number of similar consecutive units can be approximated by a symmetrical logarithmic transformation of the maximum distance between two positions in

[1] See Eq. 1.7 on page 19.

either of the dimensions (n-distance), d.

$$d = -\log_2(\max|USM_{unidirectional}|) \tag{3.3}$$

Since the USM coordinates include two CGR iterations per dimension, one forward and another backward, two distances can be extracted. The first $1, \ldots, n$ coordinates define a forward similarity estimate, d_f, and the second $n+1, \ldots, 2n$ coordinates can be used to estimate backward similarity, d_b. The former measures similarity with regard to the units preceding the one being compared and the latter does the same for those succeeding that same units. Therefore, the forward and backward distances between the positions i and j of two sequences, a and b, with a length of k_a and k_b, respectively, would be calculated as described by Eq. 3.4, defining two rectangular matrices, d_f and d_b, of size k_a x k_b (Fig. 3.2a,b).

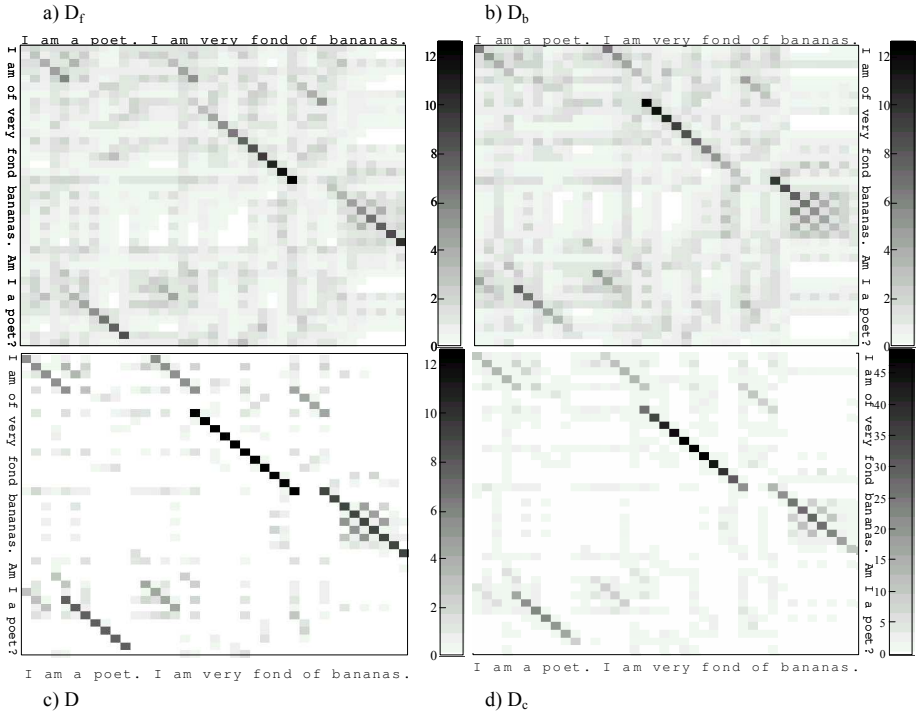


Figure 3.2: Cross-tabulation of similarity between positions of the two stanzas. The figures can be reproduced using accompanying web based USM tool (see Abstract for URL address, test data also included). a) forward distance, d_f (Eq. 3.4); b) backward distance, d_b (Eq. 3.4); c) bi-directional similarity, D, compensated for $P_3 = 0.5$, $n = 4.25$ (Eq. 3.11). Notice that the values of diagonals between similar segments estimate the number of units in the segments, although each D value is computed solely from a single pairwise comparison of USM coordinates; d) Compounded similarity, D_c, with a maximum for the mid-position of the similar segments (Eq. 3.12).

$$\begin{aligned} d_f(a_i, b_j) &= -\log_2(\max|USMb^{(j)}_{1,\ldots,n} - USMa^{(i)}_{1,\ldots,n}|) \\ d_b(a_i, b_j) &= -\log_2(\max|USMb^{(j)}_{n+1,\ldots,2n} - USMa^{(i)}_{n+1,\ldots,2n}|) \end{aligned} \tag{3.4}$$

However, the values of d necessarily overestimate the number of similar contiguous units preceding (d_f, illustration for stanza comparison in Fig. 3.2a) or succeeding (d_b, illustration for stanza comparison in Fig. 3.2b) the positions being compared. The value of d would be the exact number of contiguous similar units, h, if the starting positions for the similar segments were at a n-distance of 1, e.g. if they were in different corners of the unit hyper-dimensional USM cube. Since the initial distance is always somewhat smaller, the homology, h, measured as the number of consecutive similar units, will be smaller than d (Eq.3.5).

$$d = h + \phi, \quad \phi \geq 0 \tag{3.5}$$

The contribution of ϕ to the similarity distance, d, can be estimated from the distribution of positions in the USM map of a random sequence. A uniformly random sequence (Tino, 1999; Oliver et al., 1993; Mata-Toledo and Willis, 1997) will occupy the USM space uniformly, and, for that matter, so will the random seed of forward and backward iterative mapping, respectively Eq. 3.1 and 3.2. Therefore, a uniform distribution is an appropriate starting point to estimate the effect of ϕ , the over-determination of h by d (Eq. 3.5). Accordingly, for a given $x \in [0, 1]$, the probability, P_0, that any two coordinates, x_1 and x_2, are located within a radius $r \in (0, 1)$ is given by Eq. 3.6.

$$P_0(r) = P_0(\triangle x < r) = r(2 - r), \quad r \in (0, 1) \tag{3.6}$$

Since $P_o(r)$ is the probability of two points chosen randomly from a uniform distribution $Unif([0, 1])$ being at a distance less than r from each other, for any set of n coordinates in the USM, the likelihood of finding another position within a block distance of r would be described by raising Eq. 3.6 to the n exponent. Finally, recalling from Eq. 3.3 that sequence similarity can be obtained by a logarithmic transformation of r, the probability that the unidirectional coordinates of two random sequences are at a similar length $d > \phi$ is described by Eq. 3.7. The simplicity of the expansion for higher dimensions highlights the order-statistics properties (Arnold et al., 1992) of the n-metric introduced above (Eq. 3.3). It is noteworthy that the model for the likelihood of over-determination is the null-model, e.g. the comparison of actual sequences is evaluated against the hypothesis that the similarity observed happened by chance alone.

$$P_1(\phi) = P_1\left(\phi \geq -\log_2\left(\max(r)\right)\right) = (2^{1-\phi} - 2^{-2\phi})^n, \quad r \in (0, 1)^n \tag{3.7}$$

Finally, it is also relevant to recall that the null model for d (Eq. 3.7 for unidirectional comparisons, bi-directional null models are derived below) allows the generalization for non-integer dimensions. For example, the 19 unique unites found in the two stanzas (Tab. 3.1), define forward and backward USM maps in 5 dimensions each. However the 5th dimension is not fully utilized, as that would require $2^5 = 32$ unique units. Therefore, if there is no requirement for an integer result, the effective value of n for the two stanzas can be refined as being $n = \log_2(19) = 4.25$.

An estimation of bi-directional similarity will now be introduced that adds the forward and backward distance measures d_f and d_b. The motivation for this new estimate is the the determination of the similar length of the entire similar segment between two sequences solely by comparing any two homologous units. Accordingly, since d_f is an estimate of preceding similarity and d_b provides the succeeding similarity equivalent the sum of the two similar distances, D, (Eq. 3.8) will estimate of the bi-directional similarity, e.g. the length of the similar segment, H.

$$D = d_f + d_b = H + \phi, \quad \phi \geq 0 \tag{3.8}$$

As illustrated later in the implementation, for pairwise comparisons of homologous units of similar segments, all values of D and, consequently, of ϕ, are exactly the same. This result could possibly have been anticipated from the preceding work (Almeida et al., 2001) by noting that the value of d between two adjacent homologous units differs exactly by one unit. However, this result was in fact a surprise and one with far reaching fundamental and practical implications.

Similarly to unidirectional similarity estimation, d, the bi-directional estimate, D, being the sum of two overestimates, is also overestimated by a quantity to be defined, ϕ (Eq. 3.8). The derivation of an expression for the bi-directional overestimation will require the decomposition of P_1 (Eq. 3.7) for two cases, comparison between unidirectional coordinates of similar quadrants, P_{1a}, and of opposite quadrants, P_{1b}, as described in Eq. 3.9. Recalling from Eq. 3.2, positions in the same quadrant correspond to sequence units with the same identity, and positions in opposite quadrants correspond to comparison between coordinates of units with a different identity.

$$\begin{aligned}
P_1(\phi, n) &= \left(\frac{P_{1a}(\phi) + P_{1b}(\phi)}{2} \right)^n \Leftrightarrow \text{Eq. 3.7} \\
P_{1a}(\phi) &= \begin{cases} 1 & \text{if} \quad \phi < 1 \\ \left(2^{2-\phi} - 2^{2-2\phi}\right) & \text{otherwise} \end{cases} \\
P_{1b}(\phi) &= \begin{cases} 2^{2-\phi} - 2^{1-2\phi} & \text{if} \quad \phi < 1 \\ 2^{1-\phi} & \text{otherwise} \end{cases}
\end{aligned} \tag{3.9}$$

The need for the distinction between same and opposite quadrant comparison, which is to say between similar and between dissimilar sequence units, is caused by the fact that same quadrant comparisons are more likely to lead to higher values of d. As illustrated above for the 16th unit of the first stanza, the forward and backward coordinates must fall in the same quadrant. Consequently, the similar pattern of same and opposite quadrant comparisons for each dimension will be reflected as a bias in the bi-directional overestimation. The determination of probability, P_2, of over-determination between sums of independent unidirectional similarity estimates is derived in Eq. 3.10.

$$P_2(\phi, n) = 1 - \int_0^{\phi} (1 - P_1(\phi - \gamma)) \cdot \left(-\frac{dP_1(\gamma, n)}{d\gamma} \right) d\gamma \tag{3.10}$$

The probability of bi-directional over-determination, can now be established by using the same and opposite unidirectional comparison expressions presented in Eq. 3.9. The resulting expression for similarity over-determination by the distance between bi-directional USM coordinates, P_3, is presented in Eq. 3.11.

$$
\begin{aligned}
P_3(\phi, n) &= \left(\frac{P_{3a}(\phi) + P_{3b}(\phi)}{2} \right)^n \neq P_2(\phi, n) \\
P_{3a}(\phi) &= 1 - \int_0^{\phi} (1 - P_{1a}(\phi - \gamma)) \cdot \left(-\frac{dP_{1a}(\gamma, n)}{d\gamma} \right) d\gamma \\
P_{3b}(\phi) &= 1 - \int_0^{\phi} (1 - P_{1b}(\phi - \gamma)) \cdot \left(-\frac{dP_{1b}(\gamma, n)}{d\gamma} \right) d\gamma
\end{aligned}
\tag{3.11}
$$

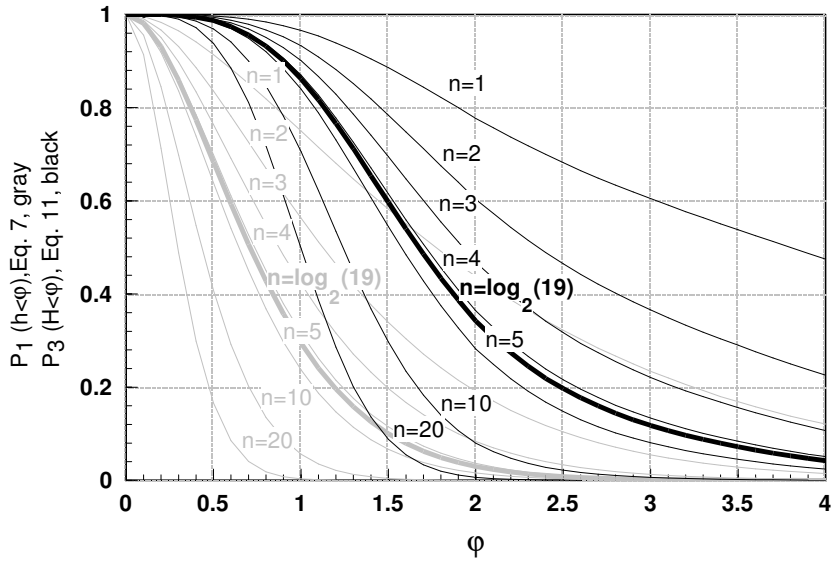


Figure 3.3: Probability distribution of similarity estimates for the uniformly random sequence null model – e.g. experimental values deviating from this model would indicate real homology, as in Fig. 3.4. The dark lines represent the numerical solution for the bi-directional over-determination, P_3 (Eq. 3.11), for different dimensionalities, n, identified by numbers in the plot. The gray lines represent the numerical solution for the same values of n, for the uni-directional over-determination, P_1 (Eq. 3.9). The solution for the dimensionality of the two stanzas, $n = \log_2(19) = 4.25$, is highlighted by a thick line, for both P_3 (thick dark line) and P_1 (thick gray line).

In Figure 3.3, the probability distribution for both unidirectional (P_1, in gray) and bi-directional (P_3, in black) comparisons is represented for different dimensions, n. It is clearly apparent that the over-determination becomes much less significant as dimensionality increases. From a practical point of view, the over-determination is of little consequence because the computational load of comparing sequences corresponds mostly to the identification of candidate pairing combinations. The fact that the n-metric unidirectional distances, d_f and d_b, defined in Eq. 3.4, and bi-directional D, defined in Eq. 3.8, are over-determined implies that the identification of similar segments between two sequences will

include false positives but will not generate false negatives. The false positive identifications can be readily recognized by comparing the sequences extracted from the coordinates, as demonstrated above for the 16th unit of the first stanza. Nevertheless, since over-determination will necessarily occur, its probability distribution was identified (Eq. 3.11, Fig. 3.3). This can also be achieved for individual values by solving Eq. 3.11 for the value of observed. For example, for the conditions of the two stanzas, the value of $\phi_{p1} = 0.5$, $n = 4.25$ is 0.71 sequence units, which is the expected median unidirectional over-determination, P_1, of d_f and d_b (Eqs. 3.5, 3.7). The corresponding probability of bi-directional overdetermination, P_3, should be somewhat above twice that value. Using Eq. 3.11, the value obtained is 1.67 similar units. Finally, it is worthy to stress that the expressions for calculation of likelihood of arbitrary levels of over-determination (Eq. 3.5–3.11) can be inverted to anticipate the level of over-determination for arbitrary probability levels. This use of the null random model is also included in the accompanying online tool (see Abstract for URL).

3.3 Discussion

H is the number of contiguous units that are similar between the two sequences aligned at the positions being compared (Eq. 3.8). This value is estimated by D, which is the sum of the overestimated number of preceding, d_f, and succeeding, d_b, homologous units (Eq. 3.4, 3.5 and 3.8). The determination of these similarity estimates, d_f and d_b, was illustrated for the two stanzas in Fig. 3.2a,b. The same values compensated for over-determination at $P_3 = 0.5$ are represented in Fig. 3.2c. The striking property of bi-directional similarity (H, Eq. 3.8) is that the D values obtained for any two homologous pair from similar segments are exactly the same. That value is an estimator of the length of the entire similar segment, H (Eq. 3.11). This is further illustrated in Fig. 3.5 for comparison of genomic sequences, where it is also observed that the values of the distances between similar segments are constant and estimate the similar length. This was a somewhat unexpected property of enormous practical value since the length of the similar segment can be determined by a single pairwise comparison between any of analogous positions. Consequently, when comparing two sequences of length k_a and k_b to identify all similar segments of length w or above, $k_a k_b / w$ pairwise comparisons will suffice. In addition, each pairwise comparison is now achievable with a single algebraic operation (Eq. 3.8) rather than requiring the conventional dynamic programming approach (Durbin et al., 1998). The computational effort of positioning database sequences in the USM state space occurs at the level of database indexing. Consequently, search algorithms based on the USM state space representation will necessarily lead to speedier implementations. In order to facilitate the comparison with dynamic programming, the software library of functions, in MATLAB format, Mathworks Inc., for the determination of USM coordinates is also provided (http://bioinformatics.musc.edu/∼jonas/usm/).

Additional measures of similarity can be derived for specific practical purposes using bi-directional and unidirectional d values. For example, the use of docking algorithms to align sequences would benefit from a measure with a

maximum value in the center of the similar segments. This could be provided by defining a compounded similarity measure, H_c, as suggested in Eq. 3.12. The behavior of H_c, which would be obtained by the overestimated value of D_c, is illustrated for the two test stanzas in Fig. 3.2d.

$$\begin{aligned} H_c &= h_f \cdot h_b \\ D_c &= d_f \cdot d_b + \phi \end{aligned} \tag{3.12}$$

The detection of similar segments in arbitrary sequences using D becomes very effective as the length of the similar segment increases. This was clear in the distribution of over-determination in Fig. 3.3 but it is even more so when the distances between sequences with homologous segments are represented.

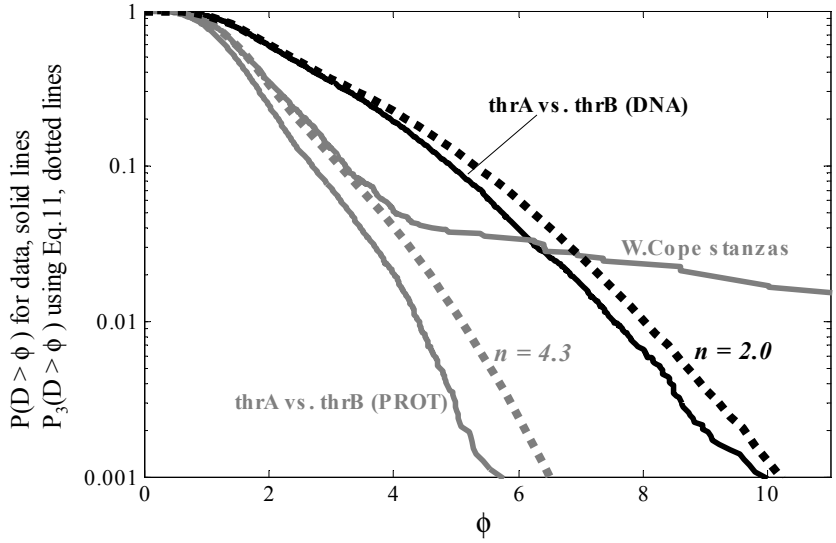


Figure 3.4: Cumulative distribution of bi-directional similarity, D, between the two stanzas and comparison of genomic and proteomic sequences of E. coli threonine gene A, thrA (2463 base pairs for the genomic sequence and 820 aminoacids for the proteomic sequence), with B, thrB (933 base pairs for the genomic sequence and 310 aminoacids for the proteomic sequence). The null model expectation, that of uniform random distribution of units, is represented by dashed lines, obtained using Eq. 3.11. for $n = 2$ (half dimensionality of USM state space for DNA) and $n = 4.3$ (half dimensionality of USM state space for proteins, $n = 4.32$, and for the two stanzas, $n = 4.25$). The solid lines represent the actual cumulative distribution of D values.

In Figure 3.4 the distances between the two stanzas are represented alongside the distances to be expected if no homology existed, apart from the coincidental (random null model, using Eq. 3.11). It can be observed for the comparison of the two stanzas (Fig. 3.4, gray lines) that H values above 4 units occur with higher frequency than allowed by the random distribution model, reflecting the presence of real homologous segments (similar words).

USM of biological sequences

The representation of biological information as discrete sequences is dominated by the fact that genomes are sequences of discrete units and so are

the products of its transcription and translation. However, not all biological sequences are composed of units that are functionally equally distinct from each other, as is the case of proteomic data and Multi-Locus Sequence Typing (MLST), (Enright et al., 2000). To avoid the issue of unit inequality and highlight the general applicability of the USM procedure, stanzas of a poem were used to illustrate the implementation instead. Nevertheless the original motivation of analyzing biological sequences is now recalled.

In the preceding report the authors have illustrated the properties of unidirectional n-metric estimation of similarity for the threonine operon of E. coli (Almeida et al., 2001). The same two two regions of thrA and thrB sequences of E.coli K-12 MG1655 are compared in Fig. 3.5 to highlight the advancement achieved by USM. It should be recalled that the particular dimensionality of DNA sequences, $n = 2$, allows a very convenient unidirectional bi-dimensional representation, which is in fact the Chaos Game Representation procedure (CGR) (Jeffrey, 1990). Consequently, CGR is a particular case of USM, obtained when $n = 2$ and only the forward coordinates are determined. This can also be verified by comparing Fig. 3.5a with a similar representation reported before (Almeida et al., 2001, Fig. 10), obtained with the same data using CGR. The advantageous properties of full (bi-directional) USM become apparent when Fig. 3.5a is compared with Fig. 3.5b. It is clearly apparent for bi-directional USM (Fig. 3.5b) that all pairwise comparisons of units of identical segments now have the same D values. This converts any individual homologous pairwise comparison into an estimation of the length of the entire similar segment. The conservation of statistical properties by the distances obtained, D, can also be confirmed by comparing observed values with the corresponding null models (Fig. 3.4). For the analysis of this figure it is noteworthy to recall that the statistical properties of prokaryote DNA are often undistinguishable from uniform randomness (Almeida et al., 2001; Román-Roldán et al., 1994; Oliver et al., 1993). The genomic sequence of the first gene of the threonine operon of E. coli, thrA, is compared with that of the second, thrB. The distribution of the resulting D values is represented in Fig. 3.4 (solid black line), alongside with the null model for that dimensionality (Eq. 3.11, with $n = \log_2(4) = 2$, gray dotted line). The genomic sequences of thrA and thrB were translated into proteomic sequences using SwissProt's online translator, applied to the 5'-3' first frame (http://www.expasy.ch/tools/dna.html). Similarly, the distribution of D values for the comparison of the proteomic thrA and thrB sequences is also represented in Fig. 3.4, alongside with the null model, Eq.3.11, for its dimensionality ($n = \log 2(uu = 20$ possible aminoacids$) = 4.32$), which is graphically nearly undistinguishable from that of the comparison between the stanzas, with $n = \log_2(uu = 19$ possible letters$) = 4.25$ (dotted gray line for the rounded value, n=4.3). Both the genomic and the proteomic distribution of D values is observed to be contained by the null model, unlike the comparison between the stanzas discussed above, where the existence of structure is clearly reflected by its distribution. The genomic and proteomic of thrA and thrB, used to illustrate this discussion, are provided with the web-based implementation of USM (see Methods for URL).

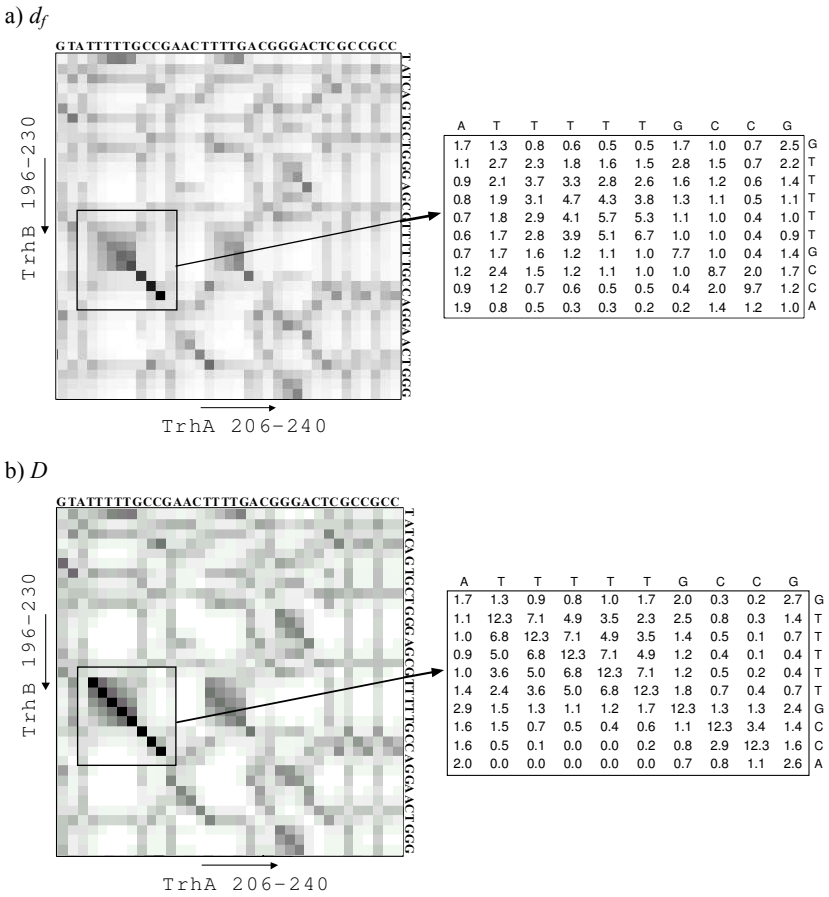


Figure 3.5: Comparison of uni-directional and bi-directional USM implementation for DNA sequences. The similarity matrices for, respectively, d_f and D values between two portions of E. coli K-12 MG1655 threonine gene A (thrA, genome positions 337–2799) and threonine gene B (thrB, genome positions 2801–3733) are presented. The numbers in the axis identify the position in the gene. Actual values of d_f and D are shown for the framed region on the table to the right. a) The d_f values were obtained by a unidirectional implementation of the USM procedure (Eq. 3.4). By comparing this figure with a similar analysis reported previously (Almeida et al., 2001, Fig. 10) for the same sequences, it can be seen that they are nearly indistinguishable, even if the exact values vary. The equivalence between unidirectional USM for $n = 2$ and CGR highlights the property that CGR is a special case of USM. The fact that the latter can be implemented for any value of n or any number of unique units justifies the universal naming; b) In this plot the same sequences were compared using bi-directional USM, and, accordingly, generate a matrix of D values (Eq. 3.8, 3.11). It is clearly apparent, and as already noted for Fig. 3.2, that D-similarity between any two homologous units is an estimate of the length of the entire homologous segment.

3.4 Conclusions

The mounting quantity and complexity of biological sequence data being produced (Roos, 2001) commands the investigation of new approaches to sequence analysis. In particular, the need for scale independent methodologies becomes even more necessary as the limitations of conventional Markov chains are increasingly noted (Hill and Singh, 1997). These limitations are bound to become overwhelming when signals such as succession schemes of the expression of over 30,000 human genes (Venter et al., 2001) become available. This particular signal would be conveniently packaged within a 30 dimension USM unit block ($n = \lceil \log_2(3.103) \rceil = 15$).

In addition, the advances in statistical mechanics for the study of complex systems, particularly in non-linear dynamics, have not been fully utilizable for the analysis of sequences due to the missing formal link between discrete sequences and trajectories in continuous spaces. The properties of USM reported above suggest that this may indeed be such a bridge. For example, the embedding of dimensions, a technique at the foundations of many time series analysis techniques offers a good example of the completeness of USM representation of sequences. By embedding the forward and backward coordinates separately, at the relevant memory length, the resulting embedded USM is exactly what would be obtained by applying USM technique to the embedded dimeric sequence itself.

3.5 Methods

3.5.1 Computation

The algorithms described in this manuscript were coded using MATLAB™ language (version 6.0 – release 12), licensed by The MathWorks Inc[2]. An internet interface was also developed to make them freely accessible through user-friendly web-pages (http://bioinformatics.musc.edu/∼jonas/usm/).

3.5.2 Source code

In order to facilitate the development of sequence analysis applications based on the USM state space, the software library of functions written to calculate the USM coordinates is provided with the web-based implementation (see address above). The code is provided in MATLAB format, which is general enough so as to be easily ported into other environments. These functions process sequences provided as text files in FASTA format. In addition to the functions, the test datasets and a brief readme.txt documentation file are also included.

3.5.3 Test data

The USM mapping proposed is applicable to any discrete sequence, even if the primary goal is the analysis of biological sequences. For ease of illustra-

[2] http://www.mathworks.com

tion and to emphasize USM's general validity, the test dataset used to describe implementation of the algorithm consists of two stanzas of a Poem by Wendy Cope, "The Uncertainty of the Poet" (Cope, 1992). In the Discussion section, USM was also applied to the DNA sequence of the threonine operon of Escherichia coli K-12 MG1655, obtained from the University of Winsconsin E. coli Genome Project (http://www.genetics.wisc.edu), and to its 5'-3' first frame proteomic translation obtained by using SwissProt online translator (http://www.expasy.ch/tools/dna.html). The three test sequence datasets are also included in the web-based USM application.

3.6 Acknowledgements

The authors thank Dr Santosh Mishra, Eli Lilly Co., for the insightful suggestions about the applicability of USM, and John H. Schwacke, at the Department of Biometry and Epidemiology of the Medical University of South Carolina for revising the coherence of mathematical deduction. The authors thankfully acknowledge financial support by grant SFRH/BD/3134/2000 to S. Vinga and project SAPIENS-34794/99 of Fundação para a Ciência e Tecnologia of the Portuguese Ministry of Science and Technology.

3.7 References

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., and Fletcher, M. (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics*, 17(5):429–437. 66, 68, 69, 70, 73, 77, 78

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. (1992). *A first course in order statistics*. Wiley, New York. 72

Basu, S., Pan, A., Dutta, C., and Das, J. (1997). Chaos game representation of proteins. *J Mol Graph Model*, 15(5):279–89. 67

Cope, W. (1992). *Serious concerns*. Faber and Faber. 67, 80

Deschavanne, P., Giron, A., Vilain, J., Fagot, G., and Fertil, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*, 16(10):1391–1399. 66

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press. 66, 75

Enright, M. C., Knox, K., Griffiths, D., Crook, D. W., and Spratt, B. G. (2000). Molecular typing of bacteria directly from cerebrospinal fluid. *Eur J Clin Microbiol Infect Dis*, 19(8):627–30. 66, 77

Fiser, A., Tusnády, G. E., and Simon, I. (1994). Chaos game representation of protein structures. *J Mol Graph*, 12(4):302–4. 66, 67

Forte, B., Mendivil, F., and Vrscay, E. R. (1998). "Chaos games" for iterated function systems with grey level maps. *SIAM Journal on Mathematical Analysis*, 29(4):878–890. 66

Goldman, N. (1993). Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res*, 21(10):2487–2491. 66, 69

Hill, K. A. and Singh, S. M. (1997). The evolution of species-type specificity in the global DNA sequence organization of mitochondrial genomes. *Genome*, 40(3):342–56. 66, 79

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res*, 18(8):2163–2170. 66, 68, 77

Mata-Toledo, R. A. and Willis, M. A. (1997). Visualization of random sequences using the chaos game algorithm. *J. Systems Software*, 39(1):3–6. 72

Nady, A. (1994). Recent investigations into global characteristics of long DNA sequences. *Indian Journal of Biochemistry and Biophysics*, 31:149–155. 66

Oliver, J., Bernaola-Galvan, P., Guerrero-Garcia, J., and Román-Roldán, R. (1993). Entropic profiles of DNA sequences through chaos-game-derived images. *J Theor Biol*, 160(4):457–70. 70, 72, 77

Pleissner, K. P., Wernisch, L., Oswald, H., and Fleck, E. (1997). Representation of amino acid sequences as two-dimensional point patterns. *Electrophoresis*, 18(15):2709–13. 67

Román-Roldán, R., Bernaola-Galván, P., and Oliver, J. (1994). Entropic feature for sequence pattern through iterated function systems. *Pattern Recognition Letters*, 15:567–573. 70, 77

Román-Roldán, R., Galván, P. B., and Oliver, J. (1996). Application of information theory to DNA-sequence analysis – a review. *Pattern Recognition*, 29(7):1187–1194. 66

Roos, D. S. (2001). Computational biology: Bioinformatics–trying to swim in a sea of data. *Science*, 291(5507):1260–1. 79

Roy, A., Raychaudhury, C., and Nandy, A. (1998). Novel techniques of graphical representation and analysis of DNA sequences – a review. *Journal of Biosciences*, 23(1):55–71. 66

Solovyev, V., Korolev, S., and Lim, H. A. (1993). A new approach for the classification of functional regions of DNA sequences based on fractal representation. *Intl J. Genomic Res.*, 1(1):109–128. 67

Tino, P. (1999). Spatial representation of symbolic sequences through iterative function systems. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 29(4):386–392. 66, 67, 72

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–51. 79

Chapter 4

Comparative evaluation of word composition distances for the recognition of SCOP relationships

Published in: Vinga, S., Gouveia-Oliveira, R. and Almeida, JS. (2004). Comparative evaluation of word composition distances for the recognition of SCOP relationships. Bioinformatics *20:2, 206–215.*

Motivation Alignment-free metrics were recently reviewed by the authors, but have not until now been object of a comparative study. This paper compares the classification accuracy of word composition metrics therein reviewed. It also presents a new distance definition between protein sequences, the W-metric, which bridges between alignment metrics, such as scores produced by the Smith-Waterman algorithm, and methods based solely in L-tuple composition, such as Euclidean distance and information content.

Results The comparative study reported here used the SCOP/ASTRAL protein secondary structure hierarchical database and accessed the discriminant value of alternative sequence dissimilarity measures by calculating Areas Under the Receiver Operating Characteristic Curves (ROC). Although alignment methods resulted in very good classification accuracy at family and superfamily levels, alignment-free distances, in particular standard Euclidean distance, are as good as alignment algorithms when sequence similarity is smaller, such as for recognition of fold or class relationships. This observation justifies its advantageous use to pre-filter homologous proteins since word statistics techniques are computed much faster than the alignment methods.

Availability All Matlab code used to generate the data is available upon request to the authors. Additional material available at http://bioinformatics.musc.edu/wmetric.

4.1 Introduction

Bioinformatics applications rely heavily on sequence comparison techniques, from searching a database with a query DNA sequence to the classification of protein domains. In most cases alignments are performed between the target sequences and the resulting alignment scores are used to calculate a measure of dissimilarity. In protein comparison, the scoring methods depend on aminoacid mutation rate information, represented as scoring matrices, and find optimal alignments between sequences by dynamic programming techniques. Alignment scores are particularly useful when sequences are known to be closely homologous since the more conserved regions are automatically detected. However, for remote homologues this approach tends to fail: proteins with less than 20% identity, a region sometimes referred to as the 'twilight zone', are not satisfactorily aligned neither its similarity detected (Pearson, 2000). It is also noteworthy that dynamic programming is computationally intensive and consequently unpractical for querying large datasets, which forces the use of some heuristics to reduce the running times, as exemplified by BLAST.

In a recent paper (Vinga and Almeida, 2003) the authors reviewed alignment-free methods for sequence comparison but did not compare them quantitatively. In that review metrics based on L-tuple composition, the focus of this report, emerged as the alignment-free technique most often proposed by other researchers. In these algorithms each sequence is mapped onto an n-dimensional vector according to its word composition. Linear algebra theory is further employed to define distances between sequences represented in those vector spaces, namely by using Euclidean distance and information content (see review for a full description and related references).

This report also presents a novel distance function between protein sequences, the W-metric, which tailors L-tuple composition methods with techniques based in alignment. This is accomplished by defining a function that includes both one-tuple composition information, more specifically the differences in aminoacid content between two proteins, and weights from the scoring matrices used in alignment methods. Although these two concepts are not new, their conjugation constitutes the novelty aspect of this metric. The weights correspond to the estimation of log-likelihood ratios between probabilities of symbols that best describe mutation rates in known homologous proteins, thus providing evolutionary information.

The usefulness of the L-tuple composition approach is associated with its light computational load, which makes it very useful in pre-filtering relevant sequences, and then using alignment algorithms to refine the searches. This type of heuristic approach is already used in programs like BLAST (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988). Although the solution may not be the optimal, it drastically shortens processing speed to the point that the method can be used to query large databases. However, a comparative study of the effectiveness of alignment-free sequence dissimilarity measures is, to the authors' best knowledge, absent from the literature. Consequently it is difficult to decide at what similarity level are alignment methods required. Such a comparative study of how these different metrics perform is reported

here. This is the main motivation for the present work, where alignment-free, linear algebra type methods are comparatively assessed. Some previous studies have reported comparative assessments of various methods (Brenner et al., 1998; Lindahl and Elofsson, 2000; Pearson, 1991, 1995), but not consistently for the same reference dataset. These studies showed however the importance of following an extensive protocol involving as many examples as possible in the assessment of any classification procedure. Only then is it possible to improve some heuristics commonly applied in sequence similarity searches and identify the best algorithmic choice for each problem category.

We compared L-tuple metrics with Smith-Waterman algorithm by Receiver Operating Characteristic curves (ROC) applying the algorithms to a subset of SCOP/ASTRAL database. This database constitutes the reference gold standard for protein secondary structure classification, which makes it a commonly used benchmark for protein structure prediction algorithms, a crucial field in computational biology applications. In addition it has a hierarchical organization that can be browsed to assess classification accuracy for each of its levels.

4.2 Systems and Methods

In the section below the W-metric, a novel word statistic distance between protein sequences is presented as well as additional background on alignment-free algorithms. In the subsequent sections the reference protein datasets and the methods used to compare the distance measures are described. Finally, the last two sections describe the algorithms and protocol used and its implementation.

4.2.1 Word statistics

There is a large body of literature on word statistics (Reinert et al., 2000), where sequences are interpreted as a succession of symbols and are further analyzed by first representing the frequencies of its small segments (L-tuples or words). This approach does not take into account any of the physicochemical or structural properties of the aminoacids or nucleotides. There is also an increasing number of studies focusing on distance definition in the frequency space of L-tuples. These definitions are a fundamental step for the subsequent application of exploratory analysis methods, such as cluster analysis and dimensionality reduction techniques. In a recent review (Vinga and Almeida, 2003) the authors overviewed these metrics and their application to biological sequences, both DNA and proteins. That review will be used as the main reference for description of the L-tuple distances and alignment-free algorithms that will be tested here. A protein X of length n is a sequence of symbols from the alphabet of all possible aminoacids: $X = s_1 \cdots s_n$, $s_i \in \mathcal{A} = \{\mathsf{A}, \mathsf{R}, \mathsf{N}, \mathsf{D}, \ldots, \mathsf{V}\}$. The mapping of X into the Euclidean space can be defined by representing X by its aminoacid composition in counts, c^X and frequencies, f^X (Eq. 4.1):

$$\begin{aligned} c^X &= (c^X_{\mathsf{A}}, c^X_{\mathsf{R}}, c^X_{\mathsf{N}}, c^X_{\mathsf{D}}, \ldots, c^X_{\mathsf{V}}) \\ f^X &= \frac{c^X}{n} \end{aligned} \tag{4.1}$$

For example, the peptide X = AARNNDAA is mapped onto the vectors $c^X = (4, 1, 2, 1, 0, 0, \ldots)$ and $f^X = (0.5, 0.125, 0.25, 0.125, 0, 0, \ldots)$. Instead of single aminoacid frequencies, longer fragments of length L could be considered (L-tuples) with resulting 20^L long vector of frequencies. One can further define a distance or dissimilarity measure between two proteins X and Y, $d(X, Y)$, based on their corresponding vectors f^X and f^Y.

4.2.2 W-metric definition

The novel W-metric hereby proposed to complement existing word composition methods is based on the quadratic form defined in Eq. 4.2. The distance between two proteins X and Y, $d^W(X, Y)$, is defined by their corresponding one-tuple frequencies, f^X and f^Y, weighted by matrix W below described.

$$\begin{aligned} d^W(X,Y) &= (f^X - f^Y)^T \cdot W \cdot (f^X - f^Y) \\ &= \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}} (f_i^X - f_i^Y) \cdot (f_j^X - f_j^Y) \cdot w_{ij} \end{aligned} \tag{4.2}$$

These quadratic forms play an important role in major theoretical and applied disciplines and scientific fields, from linear algebra to econometrics. In statistics they are used, for example, in parameter estimation and statistical tests (Schott, 1997). They represent a scoring between conveniently weighted vectors of observations. It is noteworthy that other L-tuple distances are also based on quadratic forms (Eq. 4.2), for example, when W is the covariance matrix of the data it represents Mahalanobis distance between the corresponding vectors and the standard Euclidean distance is obtained when taking only covariance matrix diagonal. The distance reduces to the squared Euclidean distance when W is the identity matrix.

The weight matrices W chosen in Eq. 4.2 can be rationalized as being scoring or aminoacid substitution matrices, instead of covariance-based weights as in other distances. These matrices, such as PAM - Point Accepted Mutation (Dayhoff et al., 1978) and BLOSUM - BLOcks SUbstitution Matrices (Henikoff and Henikoff, 1992), are used in alignment-based methods and estimate the log-likelihood ratios between probabilities of symbols that best describe mutation rates in known homologous proteins. In particular BLOSUMX matrix is estimated with ungapped aligned blocks of proteins sharing less than X% identity. PAMn matrices account for evolutionary change in protein sequences and its estimation is based on the construction of phylogenetic trees, which are subsequently used to create a Markov chain n-step transition matrix. This matrix is further transformed and normalized for conditional probabilities. For extensive description of this substitution matrices and some estimation examples see (Ewens and Grant, 2001), section 6.5.

The key idea of W-metric is to weight aminoacid composition differences between two sequences, $f_i^X - f_i^Y$, according to its relative conservation in proteins known to be homologous. The overall distance between two proteins will be the sum of these weighed factors. For example, if an aminoacid is highly conserved in known homologous sequences (high w_{ii}), two proteins with a very different frequency of this aminoacid should be less similar than if the aminoacids

are 'closer' to each other in that statistical sense. If the opposite occurs, i.e., if an aminoacid is known to have high mutational rates (low w_{ii}), the differences between its compositions in the two sequences being compared should be attenuated in the overall distance calculation. The same scheme applies to off-diagonal elements $w_{ij}(i \neq j)$: if there is a high mutation rate between these two aminoacids, it means that w_{ij} is higher than the corresponding weight of two aminoacids very different, so this component should be weighted more. The main idea is thus weighting aminoacid differences according to their similarity, given by known evolutionary information. The weighted metric hence includes both aminoacid composition information, like other alignment-free techniques, and conserved homology information, as used to score the conventional alignment algorithms.

Some variations of this metric were also tested, namely using several normalization procedures. It is appealing the low computational load associated with the calculation expressed in Eq. 4.2. It is not proven here, however, that the W matrix associated with mutation information is the best in discriminating classification levels. This can be further accomplished by using artificial neural networks (ANN) or other algorithms to optimize classification accuracy by finding a 'better' W weighting matrix.

4.2.3 ROC curve definition

The methods that will be used here to assess and compare the accuracy of classification schemes and prediction algorithms are based on the analysis of Receiver Operating Characteristic curves (ROC). This method goes back to signal detection and classification problems and is now widely applied in medical diagnosis studies and psychometric analysis (Egan, 1975). This approach is employed in binary classification of continuous data, usually categorized as positive (1) or negative (0) cases. The classification accuracy can be measured by plotting, for different threshold values, the number of true positives (TP), also named sensitivity or coverage, vs. false positives (FP), or (1−specificity), encountered for each threshold, properly normalized – see Eq. 4.3.

$$\begin{aligned} \text{sensitivity} &= \frac{\text{TruePositives}}{\text{Positives}} = \frac{TP}{TP+FN} \\ \text{specificity} &= \frac{\text{TrueNegatives}}{\text{Negatives}} = \frac{TN}{TN+FP} \\ 1-\text{specificity} &= \frac{FP}{TN+FP} \end{aligned} \tag{4.3}$$

A ROC curve is simply the plot of sensitivity vs. (1-specificity) for different threshold values. The area under a ROC curve (AUC) is a widely employed parameter to quantify the quality of a classificator because it is a threshold independent performance measure and is closely related to the Wilcoxon signed-rank test (Bradley, 1997). For a perfect classifier the AUC is 1 and for a random classifier the AUC is 0.5. For additional results and comprehensive discussion of AUC measure see (Bradley, 1997). The references (Baldi et al., 2000; Brenner et al., 1998; Green and Brenner, 2002) describe other possible classification accuracy measures not employed in this study.

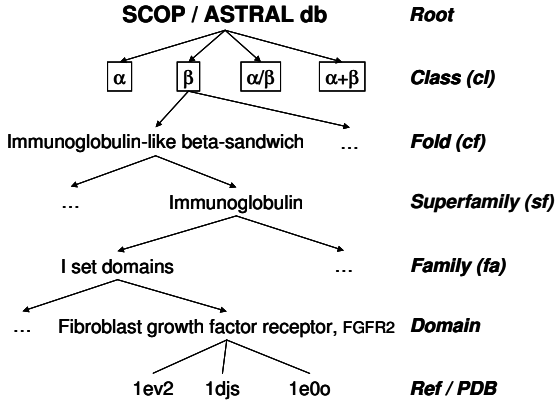


Figure 4.1: SCOP/ASTRAL db – hierarchical classification of proteins. Example of Fibroblast Growth Factor Receptor (FGFR2) classification in each of the four levels.

4.2.4 Protein test datasets – SCOP/ASTRAL classification

The sequences used to perform the tests and compare different metrics are proteins from the Structural Classification of Proteins (SCOP) database (Lo Conte et al., 2002; Murzin et al., 1995). This database consists of Protein Data Bank (PDB) entries and provides a detailed and reliable description of protein structure relationships and homology. The 3D structure analysis allows the detection of more remote homologies, since structure is typically more conserved than sequence. The fundamental unit of classification is the protein domain, which is the basic element of protein structure and evolution. The ASTRAL compendium provides additional tools and datasets (Brenner et al., 2000; Chandonia et al., 2002), namely the possibility to filter sequence sets where two different proteins have less than a chosen percentage identity to each other. This classification is a hierarchical description of proteins (see Fig. 4.1). The first two levels, family (fa) and superfamily (sf), describe evolutionary relationships; the third one, fold (cf), describes geometrical relationships or major structural similarity, and the fourth one represents protein structural class (cl). This will allow the study of each classifier for different levels of similarity.

Two different datasets were tested in order to assess the accuracy of each metric. The basic protein set, PDB40-B, was extracted directly from the ASTRAL website and corresponds to SCOP database release 1.61 (November 2002). This subset includes all the sequences that share less than 40% identity to each other and has become a benchmark test set in the evaluation of methods to detect remote protein homologies. (See, for example, (Brenner et al., 1998; Dubchak et al., 1999; Karwath and King, 2002; Lindahl and Elofsson, 2000; Luo et al., 2002; Park et al., 1997; Webb et al., 2002)) This dataset was subsequently trimmed to exclude sequences with unknown aminoacids and those belonging to families with less than 5 elements, thus obtaining the protein group named PDB40-v (see Tab. 4.1). For example, there are 232 families with only one sequence, which is not informative regarding intra-family dissimilarity,

	Classes																
	all-α				all-β				α/β				α+β				
Datasets	*do*	*fa*	*sf*	*cf*	*do*	*fa*	*sf*	*cf*	*do*	*fa*	*sf*	*cf*	*do*	*fa*	*sf*	*cf*	*Total*
PDB40-B (1.61)	**867**	409	257	151	**1051**	362	213	111	**1237**	467	190	117	**1065**	487	307	212	**4220**
PDB40-v (1.61)	**285**	35	28	27	**517**	43	30	24	**542**	58	40	31	**339**	39	37	33	**1683**
PDB40-b (1.35)	**220**	128	97	73	**309**	150	115	54	**285**	154	98	66	**240**	147	115	80	**1054**

Table 4.1: Protein datasets used in this study. For each protein set, number of sequences or domains (do), families (fa), superfamilies (sf) and folds (cf), in each class. PDB40-B: sequences that share less that 40% to each other, current release (1.61) of SCOP/ASTRAL (not tested). PDB40-v: set derived from PDB40-B (1.61) by excluding sequences with unknown aminoacids and families with less than 5 domains. PDB40-b: sequence dataset used in (Luo et al., 2002), corresponds to previous release (1.35) of the same database.

which makes these domains insufficiently representative of a family. The effect of trimming the dataset was in this way also studied. Only the four major classes were included, namely all-α class, constituted mainly by proteins with α helix; all-β class, essentially formed by β-sheet structures; α/β class, proteins with mixtures of α-helices and β-strands; and $\alpha+\beta$ class, those where α-helices and β-strands are largely segregated. Other SCOP classes include multi-domain proteins, small proteins, theoretical models and other types, and were not included in this study. See (Chothia et al., 1997) and SCOP documentation for description of protein folds and classification.

This study also considered separately another protein set from an outdated release of the SCOP database (1.35), the PDB40-b, due to the large amount of literature already published with those sequences. (See Luo et al. (2002) and corresponding references.) Table 4.1 summarizes all the sequences sets examined in this paper.

4.2.5 Protocol for comparative assessment

The comparative test procedure followed in this report was based on a binary classification of each protein pair, where 1 corresponds to the two proteins sharing the same group in SCOP database, 0 otherwise. The group can be defined at one of the 4 different levels of the database: family (fa), superfamily (sf), class fold (cf) or class (cl), exploring the hierarchical organization of the proteins in that structure. Therefore each protein pair is associated to 4 binary classifications, one for each level.

In order to compute the ROC curves, we calculated the distances between all possible protein pairs, according to the different metrics referred to and briefly described below.

The similarity measure based on alignment tested was the Smith-Waterman (SW) raw score, with no correction for statistical significance, using score matrix BLOSUM50 and a linear gaping penalty scheme, with a gap penalty of 8. The distances based on L-tuple composition evaluated were W-metric (Wm), Euclidean (eu), standard Euclidean (se), Kullback-Leibler discrepancy (ku), cosine (co), and Mahalanobis (ma). For the corresponding complete definitions and properties see (Vinga and Almeida, 2003). In W-metric calculations some

alternative weighting matrices W (Eq. 4.2) were used: these included the scoring matrices BLOSUM50, BLOSUM40, BLOSUM62 and PAM250. The following normalization procedures were also applied: take only the diagonal of W, pass all its negative values to zero, use the exponential function of the original matrix and normalize by minimum and range. However, in this printed report only the results obtained with BLOSUM50 will be presented. The variations described are documented on the online annex.

For each metric, the distances between all proteins pairs were subsequently sorted, from maximum to minimum similarity, that is, from the closest to the farthest pair. A perfect metric would completely separate negative from positive relationships, i.e., the maximum similarity would correspond always to the same group, and the binary classification obtained after this distance sorting would be the vector $(1, \ldots, 1, 1, 0, 0, \ldots, 0)$. Of course this does not happen in practice, and the classes are interspersed. The ROC curves permit to assess the level of accuracy of this separation without choosing any distance threshold for the separation point. In particular, the AUC will give us a unique number of the relative accuracy of each metric and level, according to the SCOP classification scheme. We also tested each of the four classes separately with the same procedure, to evaluate hypothetical differences between the structural classes.

4.2.6 Computation

All the algorithms were implemented in MATLAB^{TM}language (version 6 release 13). The code is available upon request to the authors.

4.3 Results and Discussion

In the following sections we present some of the results obtained. For extensive and additional results regarding all metrics and datasets see also the web page http://bioinformatics.musc.edu/wmetric, where the complete graphs and tables are shown (data not shown due to space limitations).

4.3.1 Complete dataset

ROC curves and AUC values

The Receiver Operating Characteristic (ROC) curves obtained for the complete dataset (Tab. 4.1) are presented in Fig. 4.2 (PDB40-v) and 4.3 (PDB40-b). As overviewed in the Systems and Methods section, a random classifier would have identical values of sensitivity and (1-specificity) for any threshold value considered (dashed diagonal).

Figures 4.4 and 4.5 provide graphs with the areas under ROC curves (AUC) obtained for both datasets and each SCOP level. The AUC values are typically used as a measure of overall discrimination accuracy.

As would be expected, Fig. 4.4 and 4.5 show that the AUC decreases from family to class level for both datasets. The sequence similarity between proteins sharing the same family is still well recognized. Consequently, all the distances achieve their best discrimination accuracy at this level. At class level,

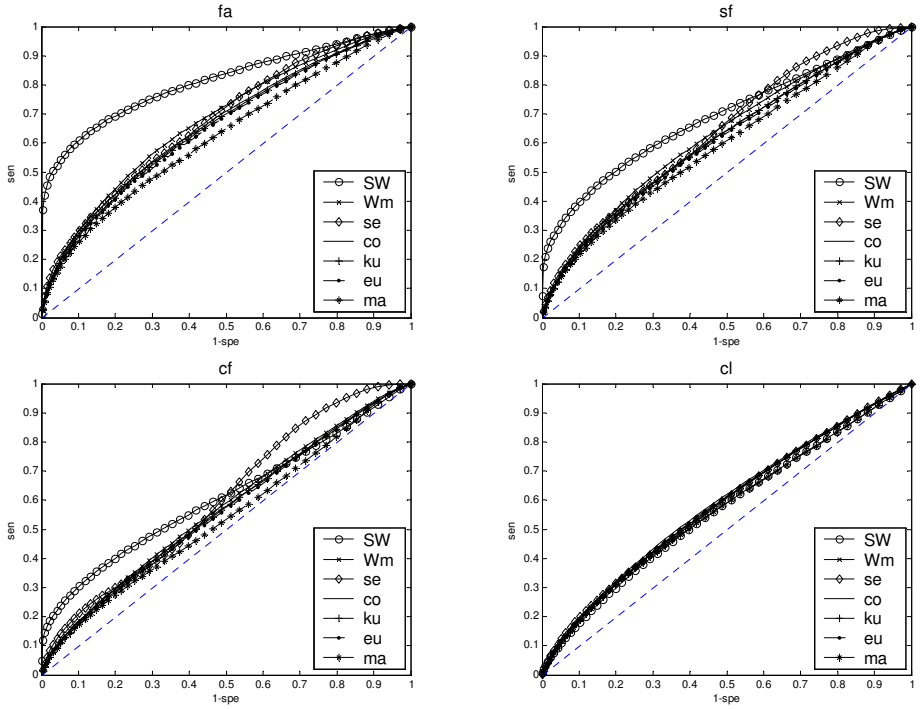


Figure 4.2: ROC curves for PDB40-v dataset. Sensitivity (sen) vs. 1-specificity (spe). SCOP levels: family (fa), superfamily (sf), class fold (cf) and class (cl). Metrics: Smith-Waterman (SW), W-metric (Wm), standard Euclidean (se), cosine (co), Kullback-Leibler (ku), Euclidean (eu) and Mahalanobis (ma). A random classifier would generate equal proportions of false positive and true positive classifications, which corresponds to the ROC diagonal (dashed line). Correspondingly, the better classification schemes have plots with higher values of sensitivity for equal values of specificity, resulting in higher values for the areas under the curve (AUC, see text). Smith-Waterman is the best at family and superfamily levels. W-metric and standard Euclidean outperform other alignment-free metrics. Standard Euclidean is the best at fold level for high sensitivity/low specificity values. For class level all metrics have similar results, slightly above random guessing.

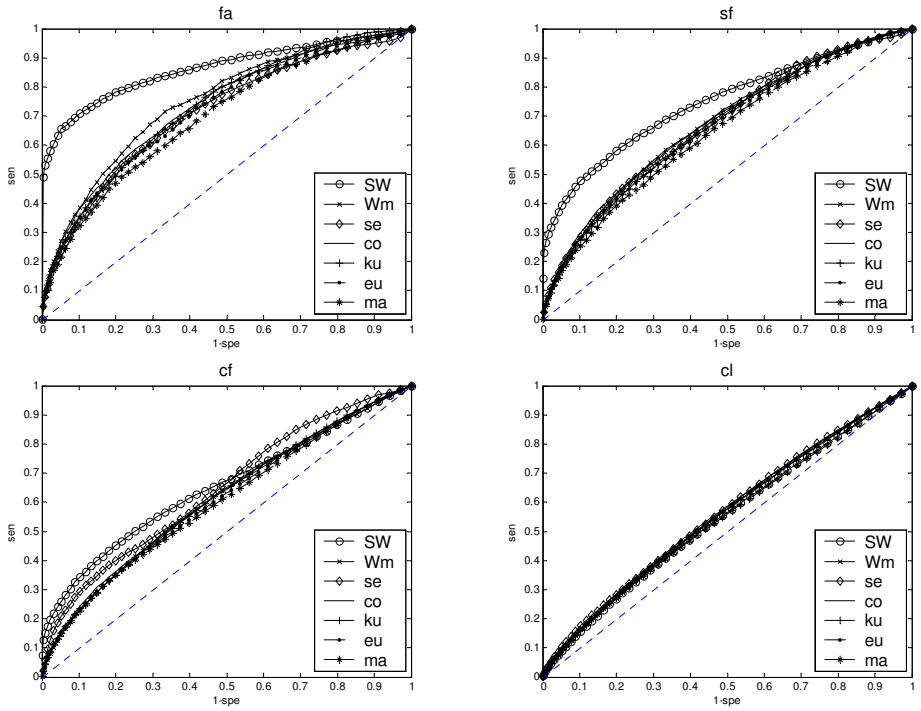


Figure 4.3: ROC curves for PDB40-b dataset. Sensitivity (sen) vs. 1-specificity (spe). SCOP levels: family (fa), superfamily (sf), class fold (cf) and class (cl). Metrics: Smith-Waterman (SW), W-metric (Wm), standard Euclidean (se), cosine (co), Kullback-Leibler (ku), Euclidean (eu) and Mahalanobis (ma). The classification accuracies for this dataset are slightly better than for the PDB40-v dataset (see Fig. 4.2). The qualitative relation between the metrics is maintained.

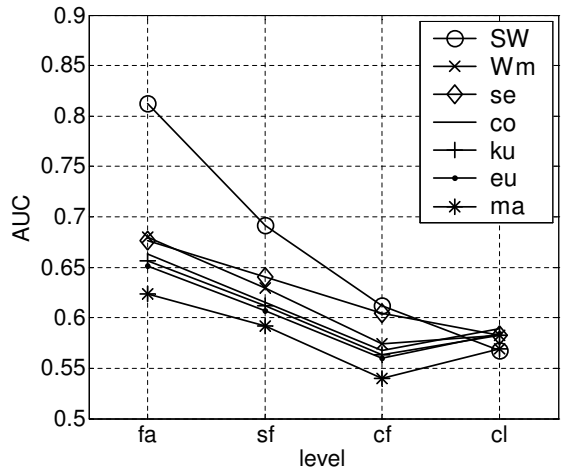


Figure 4.4: AUC values for PDB40-v dataset, for each hierarchical level. SCOP levels: family (fa), superfamily (sf), class fold (cf) and class (cl). Metrics: Smith-Waterman (SW), W-metric (Wm), standard Euclidean (se), cosine (co), Kullback-Leibler (ku), Euclidean (eu) and Mahalanobis (ma). Areas under ROC curves of Fig. 4.2. Higher AUC values correspond to better classification schemes. All the distances achieve their best discrimination accuracy at family level. This figure illustrates the loss of discrimination as the target of classification moves up in the SCOP level, from family to class.

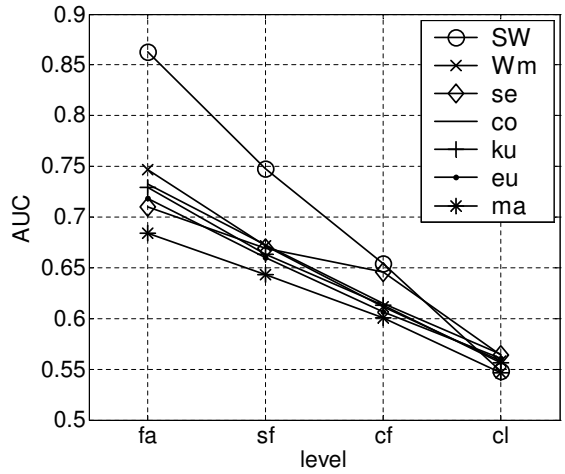


Figure 4.5: AUC values for PDB40-b dataset, for each hierarchical level. SCOP levels: family (fa), superfamily (sf), class fold (cf) and class (cl). Metrics: Smith-Waterman (SW), W-metric (Wm), standard Euclidean (se), cosine (co), Kullback-Leibler (ku), Euclidean (eu) and Mahalanobis (ma). Areas under ROC curves of Fig. 4.3. The results are slightly more discriminant for this dataset than for PDB40-v (Fig. 4.4) but with no significant changes in the metrics' relative ordering.

classification relationships reflect similar structures, which can have completely different sequences and aminoacid compositions. This underlies the observation that sequence similarity is lost, regardless of the metric, from family to class. The comparative discriminant value of the different metrics (Fig. 4.4 and 4.5) shows two clear trends. First, at family level, alignment has a clear advantage, with AUC values of 0.86 and 0.81 (PDB40b and PDB40v sets), whereas all word-statistics metrics perform at or under 0.75 and 0.68 respectively. The most discriminant word-statistics metric at family level is the novel W-metric introduced by this report (see Systems and Methods), reflecting the value of weighting the quadratic form (Eq. 4.2) by evolutionary rather than statistical criteria. At the superfamily level the advantage of alignment remains, but statistically weighting performs just as well as the W-metric. Interestingly the unweighted Euclidean metric (eu), covariance weighting (ma) and information-based Kullback-Leibler (ku) lag behind. The main surprise of this analysis is to be observed at the next level, the fold, where the standard Euclidean metric performs as well as alignment scores in both versions of SCOP, especially for the low specificity/high sensitivity range (corresponds to many False Positive relationships). In fact, standard Euclidean is clearly more discriminant than Smith-Waterman for 1-specificity values around 0.75. Finally at the class level, the absence of conserved segments in fact turns alignment into a computationally expensive procedure to score aminoacid composition differences. At this point most alignment-free metrics outperform it. The inspection of the ROC curves themselves (Fig. 4.4 and 4.5) further documents this comparison between metrics. The results obtained are slightly less discriminant for the more recent version of the protein dataset (PDB40-v) for all the levels except for class, where higher values of AUC are obtained. However, there are no significant changes in their relative ordering. It is noteworthy that there is also a dependency between levels as regards classification accuracy. Hits at a lower level may be argued to bias for more populated grouping at upper levels. However it should be noted that this study is of exploratory rather than discriminant nature, which places any pairwise comparison, regardless of the SCOP classification level, on an equal standing.

Variations in the W-metric definition

The W-metric AUC values in the previous graphics were obtained using the scoring matrix BLOSUM50. The results using BLOSUM40, BLOSUM62 and PAM250 are virtually the same and will be omitted. Nevertheless those results were compiled and are made available at the support webpage (see Availability). It is interesting to note that, although defining a different score for each domain pair, the different matrices W produce the same score ordering. Similarly, all the normalization procedures did not lead to improved discrimination, producing worse classification results but are still made available in the same webpage.

Higher order tuples

We also tested higher order word composition metrics, calculating 2 and 3-tuple distances between the domains, for Euclidean, standard Euclidean, Kullback-

Leibler and cosine. Somewhat intriguing was the fact that for all levels of classification discrimination worsened (see webpage). However it should be noted that the high dimension of the frequency vectors in these cases (respectively 400 and 8000) and the relative low dimension of the sequences length itself (mean values around 175 aminoacids), caused the frequency vector f to be very sparse. Additional problems arising from this increased dimensionality of data are the need to raise sampling size in order to maintain accuracy, which goes along with the "curse of dimensionality" (Donoho, 2000). Consequently only the results obtained for one-tuples were presented in this report. The weighting proposed, as observed before for the one-tuple scenario, might not be the best for the recognition of the relationships. One idea worth exploring would be to extract some effective higher order tuples, by adequate selection of the weights, thus optimizing the classification accuracy and hopefully avoiding the dimensionality problem. However, this would lead to discriminatory and optimization procedures, which are out of the scope of this exploratory study.

Computational performance

It is noteworthy that the Smith-Waterman algorithm is computationally intensive. Its running times can be 1000 fold longer than that of the other metrics here compared. For example in PDB40-v dataset, SW took approximately 80 hours and W-metric just 5 minutes, using a 700MHz PentiumIII with 1GB total memory. The other word composition metrics themselves have varied computation implementation efficiencies (Vinga and Almeida, 2003).

4.3.2 Stratified analysis by class

AUC values

In order to compare the metrics, we also conducted additional studies for each of the 4 classes (all-α, all-β, α/β and $\alpha+\beta$) separately. The AUC values are represented in Fig. 4.6, for Smith-Waterman alignment scores and standard Euclidean distance, the two metrics that emerged as the most discriminant in the previous analysis (Fig. 4.2–4.5) (see webpage for similar analysis for the other metrics).

It is easier to recognize family relationships by alignment (Fig. 4.6, black symbols) for proteins belonging to class all-α, where values are above the overall accuracy (AUC values ranging from 0.70 to 0.87) and for $\alpha+\beta$ class (AUC from 0.70 to 0.91). The class where these relationships seem more difficult to detect was the class all-β, where we obtained the lowest AUC values for this level (0.60 to 0.77). For superfamily level, class $\alpha+\beta$ enables a surprising accuracy for both metrics (AUC from 0.70 to 0.90) as opposed to class all-β, where the superfamily relationships are still harder to detect only by sequence inspection (AUC between 0.55 and 0.64). At fold level, all-α class retains the higher AUC values for both metrics (0.69 to 0.81). The graph obtained for PDB40-b is qualitatively the same (see webpage) with a difference: the AUC values for fold level are much lower for all-α and $\alpha+\beta$ classes for both metrics.

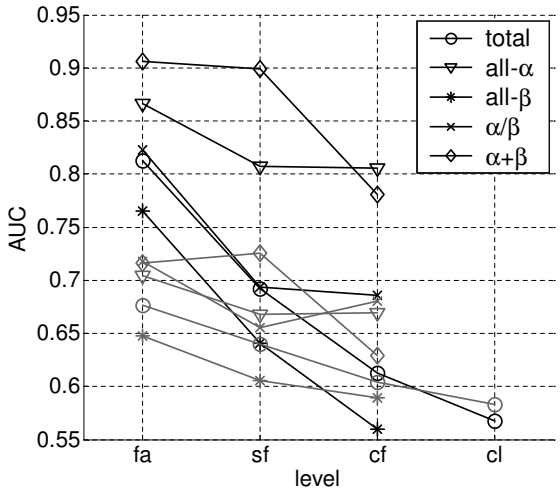


Figure 4.6: Stratified analysis by class in PDB40-v dataset. AUC values for Smith-Waterman algorithm (black) and std. Euclidean distance (gray) for each class: total set, all-α, all-β, α/β and $\alpha+\beta$. SW is generally a better classification scheme – higher AUC values. At family level the best results are for proteins belonging to classes all-α and $\alpha+\beta$; the lowest AUC values were obtained for class all-β. At superfamily level class $\alpha+\beta$ enables a surprising accuracy for both metrics as opposed to class all-β, which has the worse results. At fold level, all-α class retains the higher AUC values for both metrics.

PDB40 version datasets comparison

There is a significant improvement of discrimination accuracy for $\alpha+\beta$ class, in PDB40-v dataset. The difference in AUC values is constantly positive, for different metrics and levels, reaching a value as high as 0.21 at fold level with the Smith-Waterman alignment scores. It seems that the trimming procedure taken when obtaining PDB40-v set (see Systems and Methods) affected particularly all-α and $\alpha+\beta$ classes. It is noteworthy these quantitatively differences obtained for the two datasets.

The α-helix and β-sheet content

Judging from published reports, protein class classification is controversial. Some studies based class classification on the percentages of α-helix and β-sheets content of each chain. In a recent report a schematic table was presented with different definitions (Eisenhaber et al., 1996). As noted in that study, there are some regions of the space defined by those percentages that are not clearly classifiable. It is in this uncertainty context that SCOP offers a classification that is a global measure and takes into account all the structural information of all chains in a protein.

In order to assess the correct assignment to classes, and avoid arbitrary classification, we extracted the α and β content for each SCOP domain tested from the PDB webpage (http://www.rcsb.org/pdb/). In Figure 4.7 we present the α and β percentages for each domain, grouped by the corresponding SCOP

class classification, obtained for the PDB40-b dataset.

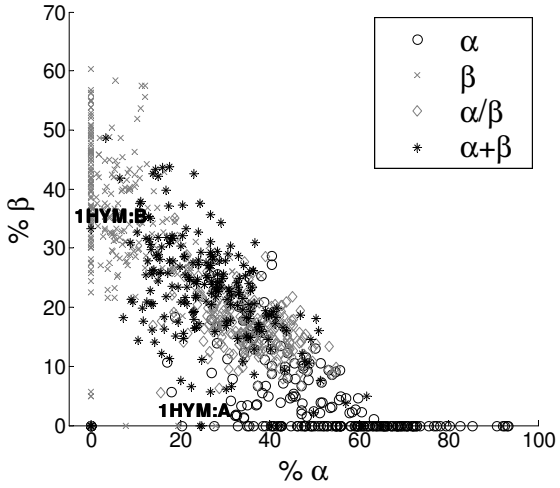


Figure 4.7: The α-helix and β-sheet content (%) for each domain in PDB40-b dataset, grouped by SCOP class. The classes are interspersed. Protein 1HYM - Trypsin inhibitor V (species: pumpkin – *Cucurbita maxima*) is globally classified in $\alpha + \beta$ class but their two chains, 1HYM:A and 1HYM:B, have contrasting α-helix and β-sheet content.

From Figure 4.7 it is apparent that some domains have arguable classifications. For example protein with PDB identification 1HYM - Trypsin inhibitor V (species: pumpkin – *Cucurbita maxima*), has two chains that correspond to two SCOP domains. Domain 1HYM:A has 24.44% of α-helix and 0% of β-sheet (labelled * symbol close to the X axis in Fig. 4.7) and domain 1HYM:B has 0%α-helix and 33.33%β-sheet (labelled * symbol close to the Y axis in Fig. 4.7). Nevertheless the whole protein was classified in the $\alpha + \beta$ class, in spite of the fact that each of its chains taken individually would be classified in other classes. The SCOP classification is global in the sense that looks to the whole protein rather than to a particular domain, therefore classifying chains of 1HYM as $\alpha + \beta$ is formally correct. Interestingly a multivariate analysis of variance (MANOVA) of the aminoacid composition in the 4 classes leads to similar results (see webpage), showing that class $\alpha + \beta$ is clearly intermixed with the others in terms of α and β content.

4.4 Conclusion

In this report we quantitatively compared several protein dissimilarity measures based on L-tuple composition with alignment scores obtained with Smith-Waterman algorithm. A new metric, the W-metric, which combines both approaches by including word statistics information weighted by scoring matrices is described.

The accuracy of each metric to detect protein relationships was assessed through the four hierarchical levels of the SCOP/ASTRAL database. The com-

parative protocol employed the areas under ROC curves (AUC), which are a good measure of overall accuracy of a classification scheme.

The Smith-Waterman alignment score was shown to be the most discriminant at family and superfamily levels. At family level, the W-metric is clearly more discriminant than the other L-tuple distances for sensitivity values between 0.5 and 0.8. From superfamily to class levels, all metrics lose discriminant power and converge to similar AUC values, which makes it counterproductive to use computational intensive alignment algorithms to detect those relationships. At fold level standard Euclidean distance outperforms most of the metrics, achieving an unexpected accuracy for high sensitivity/low specificity range. This important result anticipates its use in providing a conservative pre-screening procedure for this problem category. In fact, since L-tuple methods are computationally much lighter, they can be useful to pre-select similar proteins before applying the alignment algorithms, thus combining the powerful aspects of each technique and greatly improving heuristic methods in sequence similarity searches.

The graph showing α-helix and β-sheet content for each domain shows that class classification cannot be inferred directly from that information, at least for mixed classes. Therefore it might be advantageous in some applications to reconsider protein class classification of each domain by exploring the distribution of sequence distances by unsupervised learning algorithms.

4.5 Acknowledgements

The authors thank John Schwacke, of the Medical University of South Carolina, for providing streamlined MATLAB code for Smith-Waterman alignment and Steven Brenner of the University of California at Berkeley for precious advice in the use of the PDB40-B set. The authors thankfully acknowledge the financial support by grants SFRH/BD/3134/2000 to S.V. and SAPIENS/34794/99 from Fundação para a Ciência e a Tecnologia (FCT) of the Portuguese Ministério da Ciência e do Ensino Superior. R.G.-O. thankfully acknowledges grant QLK2-CT-2000-01020 (EURIS) from the European Commission. This work was also supported in part by the NHLBI Proteomics Initiative through contract N01-HV-28181, and a Cancer Center grant from the Department of Energy (C.E. Reed, PI).

4.6 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215:403–410. 84

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–24. 87

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159. 87

Brenner, S. E., Chothia, C., and Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA*, 95(11):6073–8. 85, 87, 88

Brenner, S. E., Koehl, P., and Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–6. 88

Chandonia, J. M., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. (2002). ASTRAL compendium enhancements. *Nucleic Acids Res*, 30(1):260–3. 88

Chothia, C., Hubbard, T., Brenner, S., Barns, H., and Murzin, A. (1997). Protein folds in the all-beta and all-alpha classes. *Annu Rev Biophys Biomol Struct*, 26:597–627. 89

Dayhoff, M. O., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In Dayhoff, M. O., editor, *Atlas of protein sequence and structure*, volume 5 - supplement 3, pages 345–352. National Biomedical Research Foundation, Washington, D.C. 86

Donoho, D. L. (2000). Aide-memoire. high-dimenional data analysis: the curses and blessings of dimensionality. 95

Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S. H. (1999). Recognition of a protein fold in the context of the structural classification of proteins (SCOP) classification. *Proteins*, 35(4):401–7. 88

Egan, J. P. (1975). *Signal detection theory and ROC-analysis*. Academic Press, New York. 87

Eisenhaber, F., Frommel, C., and Argos, P. (1996). Prediction of secondary structural content of proteins from their amino acid composition alone. ii. the paradox with secondary structural class. *Proteins*, 25(2):169–79. 96

Ewens, W. J. and Grant, G. R. (2001). *Statistical methods in bioinformatics: an introduction*. Springer, New York. 86

Green, R. E. and Brenner, S. E. (2002). Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. *Proc IEEE*, 90(12):1834–1847. 87

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, 89(22):10915–9. 86

Karwath, A. and King, R. D. (2002). Homology induction: the use of machine learning to improve sequence similarity searches. *BMC Bioinformatics*, 3(1):11. 88

Lindahl, E. and Elofsson, A. (2000). Identification of related proteins on family, superfamily and fold level. *J Mol Biol*, 295(3):613–25. 85, 88

Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*, 30(1):264–7. 88

Luo, R. Y., Feng, Z. P., and Liu, J. K. (2002). Prediction of protein structural class by amino acid and polypeptide composition. *Eur J Biochem*, 269(17):4219–25. 88, 89

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40. 88

Park, J., Teichmann, S. A., Hubbard, T., and Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J Mol Biol*, 273:349–354. 88

Pearson, W. R. (1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11(3):635–50. 85

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci*, 4(6):1145–60. 85

Pearson, W. R. (2000). Protein sequence comparison and protein evolution. *Tutorial – ISMB2000*. 84

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 85(8):2444–8. 84

Reinert, G., Schbath, S., and Waterman, M. S. (2000). Probabilistic and statistical properties of words: an overview. *J Comput Biol*, 7(1–2):1–46. 85

Schott, J. R. (1997). *Matrix analysis for statistics*. John Wiley, New York. 86

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison–a review. *Bioinformatics*, 19(4):513–523. 84, 85, 89, 95

Webb, B.-J. M., Liu, J. S., and Lawrence, C. E. (2002). BALSA: Bayesian algorithm for local sequence alignment. *Nucleic Acids Res*, 30(5):1268–1277. 88

Chapter 5

Rényi continuous entropy of DNA sequences

Published in: Vinga, S. and Almeida, JS. (2004). Rényi continuous entropy of DNA sequences. Journal of Theoretical Biology *231:3, 377–388.*
Supplementary material added: appendix D with additional deductions; description of MATLAB functions.

Entropy measures of DNA sequences estimate their randomness or, inversely, their repeatability. L-block Shannon discrete entropy accounts for the empirical distribution of all length-L words and has convergence problems for finite sequences. A new entropy measure that extends Shannon's formalism is proposed. Rényi's quadratic entropy calculated with Parzen window density estimation method applied to CGR/USM continuous maps of DNA sequences constitute a novel technique to evaluate sequence global randomness without some of the former method drawbacks.

The asymptotic behavior of this new measure was analytically deduced and the calculation of entropies for several synthetic and experimental biological sequences was performed. The results obtained were compared with the distributions of the null model of randomness obtained by simulation. The biological sequences have shown a different p-value according to the kernel resolution of Parzen's method, which might indicate an unknown level of organization of their patterns. This new technique can be very useful in the study of DNA sequence complexity and provide additional tools for DNA entropy estimation.

The main MATLAB applications developed are available at the webpage http://bioinformatics.musc.edu/renyi. Specialized functions can be obtained from the authors.

5.1 Introduction

Biological sequences are the ultimate support for the description of biological systems. However, the mathematical foundations for sequence analysis still very incompletely address issues of scale (Markov order) dependency. Consequently,

the analogy between DNA and a structured language with rules and the correspondence to coding theory has long been the object of theory and, more recently, computational exploits. The concept of entropy, as accessed by information content (Shannon) or algorithmic complexity (Kolmogorov), is of pivotal importance in the quantitative assessment of this analogy. Entropy is a measure of the degree of randomness of a system. This definition was first applied to the study of gases in thermodynamics, measuring the level of organization of the system by the number of allowable microstates. Later, Claude Shannon's pioneer paper (Shannon, 1948) founded the field of Information Theory (IT), establishing a relationship between entropy, probability of source outcomes and the ability of conveying information. IT is now an independent discipline, founded on probabilistic and axiomatic grounds (Khinchin, 1957). Alfréd Rényi further extended Shannon's entropy concepts providing more flexibility to the uncertainty measure definition (Rényi, 1961; Rényi, 1966).

As mentioned, information theory was first developed to study transmission of messages over a channel in engineering applications. Only much later IT was applied to the study of biological sequences. Over three decades ago, in a seminal book (Gatlin, 1972), Lila Gatlin explored the relation between information theory and biology and the applicability of entropy concepts to DNA sequence analysis. The most widely used definition is the L-block Shannon discrete entropy $H_{Shannon} = -\sum_i p_i \log_2 p_i$, where p_i represents the observed probabilities of words of length L (or L-tuples). Subsequently, a wide range of applications developed based on these important concepts, namely intron/exon comparison and gene prediction, the role of repeats in entropy estimates (Herzel et al., 1994b), the study of the inherent stochasticity of this quantity (Jimenez-Montano et al., 2002), Zipf and redundancy analysis (Mantegna et al., 1994) and several other applications (Chechetkin and Lobzin, 1996; Lio et al., 1996; Xiao et al., 2002).
One of the major problems when calculating Shannon's L-block entropy is the finite size sample effect (Herzel et al., 1994a), given that real biological sequences are always finite. The resulting convergence problem causes the systematic underestimation of entropy when L increases. This problem was mentioned and was partially corrected in several published studies (Schmitt and Herzel, 1997), but the main sample effect always persists for some higher word length.

Following also Shannon's pioneering work, where it was shown that there is a relation between the entropy of a source and the length of the optimal binary code transmitted, some studies have applied compression methods to DNA sequences. There is an association between these concepts: a sequence with low entropy, i.e. high redundancy, will be more compressible. Therefore, the length of the compressed sequence will give an estimate of its complexity, and consequently its entropy (Farach et al., 1995). The drawback of this method is that the compression procedures are likely to fail to recognize complex organization levels in the sequences. This is particularly relevant for biological sequences where some level of redundancy spans all scales (multiple codons for the same coded aminoacid, repeats in regions with high recombination, and finally, gene, and even genome, duplication). Several compression techniques have been developed to estimate entropy/complexity (Lanctot et al., 2000; Loewenstern and

Yianilos, 1999). One striking result obtained in several studies reports the low decrease of complexity for biological sequences. For example, it was shown that proteins are almost random polypeptides, having 99% of the complexity of random proteins (Orlov et al., 2002; Weiss et al., 2000), which shows that non-randomness is not required for proteins to be functional. Similar results were obtained for bacterial DNA, where the entropy, measured with standard methods, is almost the same as for random sequences (Almeida et al., 2001; Hariri et al., 1990; Oliver et al., 1993). This fact might be associated with the higher information holding capacity of almost random sequences: if they were perfectly random error detection would be impossible. On the other hand, if sequences were perfectly deterministic or predictable, and therefore highly redundant, no information could be stored and transmitted. These scenarios illustrate the subtle balance between error and fidelity that is apparent in biological sequences. This also establishes the need for new measures of randomness capable of exploring scales simultaneously, similarly to the Biological processes supported by the sequence themselves.

Other concepts are associated with entropy, such as the Linguistic Complexity (LC) (Crochemore and Verin, 1999; Gabrielian and Bolshoy, 1999; Troyanskaya et al., 2002) that accounts for L-tuple variability in sliding windows. An alternative format where entropy can be investigated is the fractal analysis of DNA sequence representations as random walks in several dimensions and the study of long-range correlations and scaling features (Almeida et al., 2001; Berthelsen et al., 1992; Herzel and Grosse, 1995; Stanley et al., 1999). The results obtained are comparable to other previous methods. It is worth mentioning that derived entropy concepts were also developed to classification problems, for example the Kullback-Leibler discrepancy between sequences and the Kolmogorov complexity (Sadovsky, 2003; Vinga and Almeida, 2003).

In the present work a new measure of entropy based on Rényi definition (Rényi, 1961; Rényi, 1966) is proposed. This measure is based on the Chaos Game Representation/Universal Sequence Maps (CGR/USM) (Almeida and Vinga, 2002; Jeffrey, 1990) of DNA, which maps a sequence onto a continuous space and permits the depiction of all its L-tuple frequencies in an invariant representation. This representation is closely related to the genomic signature concept (Deschavanne et al., 1999) and with fractal theory (Barnsley, 1998; Yu et al., 2004). It was proven elsewhere that these maps generalize Markov chain models on any order (Almeida et al., 2001), which can be used to extract discrete entropy measures; the present study proposes the additional extraction of continuous entropy measures from the same maps.

This representation was also shown to be preferable to variable length Markov models for the prediction of several finite memory models (Tino and Dorffner, 2001). There is also a close relation between fractal dimensions of these maps and the discrete entropies of the sequences under study (Tino, 1999, 2002), which further justifies its application in this paper. It is also noteworthy that in a very recent paper, the application of Rényi discrete entropy measure to the identification of DNA binding sites was significantly better than the Shannon-based results, which reflects the flexibility gained when using the Rényi formulation (Krishnamachari et al., 2004).

In the present report, the Parzen window density estimation of CGR/USM maps associated to Rényi entropy calculations was used, which has very significant computational advantages, described and discussed below. The association of this method with Rényi entropy estimation was already employed with success in machine learning and data mining techniques for the non-supervised classification of several classes of objects (Principe et al., 2000). These three concepts allow the calculation of a continuous entropy measure of a discrete sequence and will be used in this report to analyze both real and synthetic DNA sequences, with reference Rényi entropy values obtained by Montecarlo simulation experiments. This study follows on a review of alignment-free methods for sequence comparison (Vinga and Almeida, 2003), which was then put to use in a systematic analysis of functional protein families as described by the SCOP database (Vinga et al., 2004).

5.2 System and methods

This section will describe the new continuous entropy measure for DNA sequences, H_2 – Rényi quadratic entropy, based on Chaos Game Representation/Universal Sequence Maps – CGR/USM (Almeida and Vinga, 2002; Jeffrey, 1990) and probability density estimation (pdf) by the Parzen window method (Parzen, 1962). Some properties of H_2 are explored, namely its asymptotic behavior and the results for random sequences, and the DNA sequence dataset tested is described.

5.2.1 CGR/USM representation of a sequence

Chaos Game Representation (CGR) was first proposed nearly a decade and a half ago as a method to identify patterns in DNA sequences (Jeffrey, 1990). The algorithm is based on iterated function systems of fractal theory (Barnsley, 1998) and maps a discrete sequence of symbols onto a continuous space. The method was later extended to n-dimensional alphabets, named Universal Sequence Maps (USM) (Almeida and Vinga, 2002), thus allowing the representation of proteins and natural languages texts. The algorithm assigns each symbol of the sequence alphabet to a corner of a hypercube and represents a sequence by successively going half the distance to the corner corresponding to the following symbol in the sequence. For example the CGR mapping $x_i \in \mathbb{R}^2$ of a N-length DNA sequence $S = s_1 s_2 \dots s_N$, $s_i \in \mathcal{A} = \{\mathsf{A}, \mathsf{T}, \mathsf{C}, \mathsf{G}\}$, $i = 1, \dots, N$ is given by the following Eq. 5.1:

$$\begin{cases} x_0 \sim Unif(0,1)^2 \\ x_i = x_{i-1} + \frac{1}{2}(y_i - x_{i-1}), i = 1, \dots, N \end{cases} \quad \text{where} \quad y_i = \begin{cases} (0,0) & \text{if } s_i = \mathsf{A} \\ (0,1) & \text{if } s_i = \mathsf{C} \\ (1,0) & \text{if } s_i = \mathsf{G} \\ (1,1) & \text{if } s_i = \mathsf{T} \end{cases} \tag{5.1}$$

The starting point x_0 is randomly chosen within the square, which corresponds to the Uniform distribution in the equation above (in the original report

proposing CGR this point was set as the middle of the square). Each symbol is then mapped onto a unique point in the CGR map using Eq. 5.1. In an earlier report it was shown that CGR representation generalizes transition probability tables of Markov chains (Almeida et al., 2001). This is due to an important CGR property represented in Fig. 5.1: sub-strings with the same suffix of length p are in the same sub-quadrants of size 2^{-p}. This means that if a motif is highly repeated in the sequence, the area that corresponds to that motif is more densely populated. Accordingly, the sub-quadrants that correspond to missing motifs or sub-strings will be empty. A random sequence will fill the space uniformly. The Rényi entropy measure here proposed will be anchored in this important property. Although visually appealing as a 2D representation of

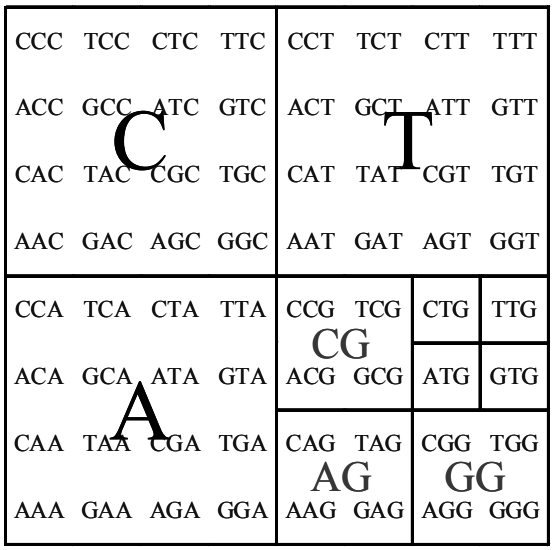


Figure 5.1: Chaos Game Representation (CGR) suffix property. Sequences ending in a specific sub-string are in the square labeled with that suffix, creating a fractal like figure. Represents equivalence of CGR and Markov chain models by transition probability matrices extraction.

DNA sequences (4-symbol alphabet), CGR considers symbol pairs differently: the Euclidean distance between the symbols represented in each diagonal (in this case symbol 'A'–'T' and 'C'–'G' – see Fig. 5.1) is $\sqrt{2}$, which is different from the distance between the other symbol-pairs, equal to the square size, 1. To avoid this bias between symbols, 4D sparse USM representation (Almeida and Vinga, 2002) is used, since all the properties are maintained, and assign each DNA symbol to the following binary numbers on Eq. 5.2, where each point representing each symbol is given by $x_i \in \mathbb{R}^4$, $i = 1, \ldots, N$ (N is the length of the sequence):

$$\begin{cases} x_0 \sim Unif(0,1)^4 \\ x_i = x_{i-1} + \frac{1}{2}(y_i - x_{i-1}), i = 1, \ldots, N \end{cases}$$

$$\text{where} \qquad y_i = \begin{cases} (1,0,0,0) & \text{if } s_i = \mathsf{A} \\ (0,1,0,0) & \text{if } s_i = \mathsf{C} \\ (0,0,1,0) & \text{if } s_i = \mathsf{G} \\ (0,0,0,1) & \text{if } s_i = \mathsf{T} \end{cases} \tag{5.2}$$

Although visualization power might be lost, a more accurate and unbiased model for sequence representation is achieved. CGR/USM representation will be used given its flexibility and because this continuous space representation of a discrete sequence allows more generalization in the tools that might be applied. The variables employed below will be the USM coordinates sample points $\{x_i\}_{i=1,\ldots,N}$.

5.2.2 Rényi continuous entropy definition

Entropy concepts were first introduced to study information transmission over a channel (Shannon, 1948), borrowing a concept first used in thermodynamics of gases. Shannon's entropy is a measure of randomness of a source based on the probabilities pi of all its possible states or outcomes $i = 1, \ldots, N$. The maximum entropy occurs when all the states have the same probability, which corresponds to the highest degree of randomness of the system. On the contrary, when $p_i = 1$ for some i, the entropy is zero, which corresponds to a deterministic system. For a comprehensive introduction to information theory and applications see (Ash, 1990; Cover and Thomas, 1991).

Later Alfréd Rényi proposed a natural extension of Shannon's entropy definition (Rényi, 1961; Rényi, 1966) for discrete and continuous probability density functions (pdf). The Rényi entropy of order $\alpha \geq 0$, $\alpha \neq 1$ of a continuous pdf $f(x)$ is defined in Eq. 5.3:

$$H_\alpha = \frac{1}{1-\alpha} \ln \int f^\alpha(x) dx \tag{5.3}$$

In this report natural logarithms will be used otherwise noticed. It can be shown that $\lim_{\alpha \to 1} H_\alpha$ is the Shannon's entropy, proving that Rényi's formulation is a generalization of the former measure. The parameter α can be interpreted as the inverse of the temperature in thermodynamic systems. When $\alpha = 0$, which corresponds to infinite temperature, the entropy attains its maximum and is simply given by the volume of the support set. It is noteworthy that entropies of continuous functions do not have the same properties as discrete probability functions, namely their positivity.

This report is focused on Rényi's quadratic entropy, $\alpha = 2$, because this leads to an important computational simplification, described in the next section, obtained for Gaussian kernels. The expression of Rényi quadratic entropy used in this report is given by Eq. 5.4.

$$H_2 = -\ln \int f^2(x) dx \tag{5.4}$$

The function $f(x)$ is the density of points of the CGR/USM maps, i.e. the continuous density of a given coordinate; x represents the CGR/USM coordinate of a symbol. Rényi entropy of a CGR map will hopefully give important clues about the randomness of the original represented sequence.

5.2.3 Parzen window density estimation

There are several non-parametric methods to perform an estimation of a continuous probability density function (pdf). These consist on estimating $\hat{f}(x)$ given a sample of N independent identically distributed random variables $a = (a_1, a_2, \ldots, a_N)$ with common underlying probability density function $f(x)$. $(a_i, i = 1, \ldots, N$ are also known as the training points.) Parzen window method (Parzen, 1962) is one of the most widely used kernel-based methods and consists on the choice of a specific weighting function or kernel $\kappa(x)$, which usually satisfies the properties of a pdf, namely $\int \kappa(x)dx = 1$. The estimation $\hat{f}(x)$ of a random vector x is a linear combination of the kernels centered in the observed sample points a_i (Eq. 5.5):[1]

$$\hat{f}(x; a) = \frac{1}{N} \sum_{i=1}^{N} \kappa(x - a_i) \tag{5.5}$$

For differentiable kernels $\kappa(x)$ this procedure corresponds to smoothing the original discrete empirical distribution while keeping all pdf properties, namely $\int \hat{f}(x; a)dx = 1$. This method is used in a wide range of applications, from neural networks to classification problems, given its flexibility and convenient properties of the estimate when the sample size tends to $+\infty$, such as consistency and asymptotic normality.

In this report the p-dimensional Gaussian or Normal kernel g_p is chosen, to take full advantage of some important properties of this distribution. This corresponds to the pdf Gaussian Eq. A.1 in appendix. A spherical symmetric Gaussian kernel will be further assumed, with mean zero $\mu = 0$ and diagonal covariance matrix $\Sigma = \sigma^2 I_p$, where I_p is the $p \times p$ identity matrix, with simplified formula given by Eq. 5.6:

$$\kappa(x) = g_p(x; 0, \sigma^2 I_p) = \frac{1}{(2\pi)^{p/2}\sigma^p} \exp\left(-\frac{1}{2\sigma^2} x^T x\right) \tag{5.6}$$

5.2.4 Simplification of Rényi entropy calculation for USM maps

The proposed entropy measure conjugates the three last concepts described. The Rényi quadratic entropy of the USM map density $f(x)$ is calculated using Parzen's method with a Gaussian kernel. An additional simplification obtained in the calculations of Rényi quadratic entropy uses another important property of Gaussian functions: the convolution of two Gaussians is also a Gaussian. This

[1] The original formulation includes another parameter, the window width h, and estimates the pdf as $\hat{f}(x; a) = \frac{1}{Nh} \sum_{i=1}^{N} \kappa(\frac{x - a_i}{h})$. In this report, the width was set to $h = 1$, meaning that only the kernel variance effect on the estimation was studied.

important simplification avoids the calculus of the integral with numeric methods as shown below. (See Appendix B for the complete deduction.) This leads to the simplified expression for the Rényi quadratic entropy given by Eq. 5.7:

$$\begin{aligned} H_2(USM) &= -\ln \int \hat{f}^2(x)dx \\ &= -\ln \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} g_4(a_i - a_j; 2\sigma^2 I_4) \\ &= -\ln \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{16\pi^2\sigma^4} \exp\left(-\frac{1}{4\sigma^2} d_{ij}\right) \end{aligned} \tag{5.7}$$

where d_{ij} is the squared Euclidean distance between the USM points (coordinates) a_i and a_j. The global Rényi quadratic entropy of the USM map depends only on all pairwise squared Euclidean distances d_{ij}, i.e. all interactions between pairs of samples, reason why some authors call this expression an information potential, as analogous to gravitational and electromagnetic fields (Principe et al., 2000), and look at these quantities as cost functions for machine learning algorithms. It should be stressed that there is no approximation in this assessment, apart from the pdf estimation.

The combination of Rényi quadratic entropies and Parzen's method has been extensively explored in classification problems and machine learning techniques (Gokcay and Principe, 2000, 2002; Principe and Xu, 1999; Principe et al., 2000), but never, to our knowledge, specifically applied to bioinformatics problems.

The study described here extends L-tuple entropy calculations to a continuous measure, hopefully extracting more features in this new USM map. Although this measure is still dependent of one parameter, the kernel variance σ^2, the protocol described below will overcome this problem by the systematically use of random and synthetic control sequences, obtained by simulation.

5.2.5 Asymptotic properties of H_2 and random sequence simulation

The study of the function H_2 theoretical limiting behavior is of critical importance since this provides a threshold for further comparison with real biological sequences. Given that the entropy of a sequence H_2 in this method depends on the variance of the Gaussian kernel used, it is necessary to study H_2 properties for different σ^2 values, namely the asymptotic behavior of $H_2 = H_2\left(f(\sigma^2)\right)$ when σ^2 tends to $+\infty$ and 0.

It is proven in this study that the graph of the function $H_2 = H_2(\ln \sigma^2)$ has two linear asymptotes of the form $H_2^+ = m \ln \sigma^2 + b$ for $\ln \sigma^2 \to +\infty$ and $H_2^- = m' \ln \sigma^2 + b'$ for $\ln \sigma^2 \to -\infty$ that are independent of the sequences under study. This means that in the graph H_2 vs. $\ln \sigma^2$ every sequence will tend to the straight asymptote lines given by the following Eq. 5.8:

$$\begin{aligned} H_2^+ &= 2 \ln \sigma^2 + \ln 16\pi^2 \\ H_2^- &= 2 \ln \sigma^2 + \ln 16\pi^2 + \ln N \end{aligned} \tag{5.8}$$

For demonstration details, see Appendix C. This result suggests the use of log graphs, i.e. $H_2 = H_2(\ln \sigma^2)$.

The asymptotic result for $\ln \sigma^2 \to +\infty$ is a delimiting reference since all sequences, independently of their length and type, will tend to this straight line when the variance σ^2 increases. The uniformity of H_2 for higher values of σ^2 can be made intuitive by noting that higher variance corresponds to wider and flatter Gaussians, and, consequently, more uniform the pdf estimation will be, independently of the sequences represented.

Analogously, when $\ln \sigma^2 \to -\infty$ the linear asymptote H_2^- has the same variance-dependent component and a term that is dependent of the sequence length ($\ln N$) but again is independent of the sequence randomness level. It is interesting to note that both asymptotes correspond exactly to the entropy of a sequence with just one symbol (single-point USM map), since in this case ($N = 1$), Eq. 5.7 reduces to $H_2(USM)|_{N=1} = 2 \ln \sigma^2 + \ln 16\pi^2 = H_2^+ = H_2^-|_{N=1}$.

Another pivotal property is the H_2 behavior for random sequences. Given the difficult mathematical deduction of the corresponding distribution of H_2, with no explicit algebraic solution, all the properties were calibrated with reference to simulation studies. Hence simulations were performed generating, for each kernel variance, 10^4 sequences of the same length as the original sequence dataset (all sequences have length $N = 2000$) and their H_2 value was calculated. With this procedure it is possible to obtain empirical distributions of the Rényi quadratic entropy values for the null model of randomness, thus permitting the subsequent comparison with the H_2 values of the sequence test set (see Results section below). This simulation study also provides standard deviation values for the measure H_2, which are important to further perform hypothesis testing and confidence intervals calculation.

The H_2 properties for random sequences of different lengths are also studied by simulation, considering again the mathematical difficulties due to the inherent stochasticity of the entropy measures. Similarly to the previous case, but with fewer replicates, 1000 random sequences are simulated with lengths varying between 1 and 4000 symbols, providing threshold values of H_2 entropies and standard deviations of this measure, again in the null random model case. These empirical values will help to model H_2 behavior as a function of the sequence length.

5.2.6 DNA sequence dataset description

This study of Rényi entropies was mostly focused on artificial DNA sequences. This ensures an accurate interpretation of the results, since it is often very difficult to have a rigorous parameter control when dealing with real biological data. Hence, several DNA sequences of different types were generated to produce calibrated values, adding in this study one biological sequence as a comparison target. The following Table 5.1 describes all the DNA sequences used, where for consistency a cutoff length of $N = 2000$ was imposed. These DNA sequences include several categories, including random sequences with equiprobability of symbols, i.e. $p_{\mathsf{A}} = p_{\mathsf{T}} = p_{\mathsf{C}} = p_{\mathsf{G}} = 0.25$, which corresponds to the absence of any structure and/or motifs (*rand*). Another type simulated

Name	*Sequence description*
rand	random
m3	random with inserted motif L=3 'ATC'
m4	random with inserted motif L=4 'ATCG'
m5	random with inserted motif L=5 'ATCGA'
m7e	random with inserted motif with error, L=7 'ATC*AGC', * denotes any symbol
R1	repeated motif L=1, 2000 times 'A'
R5	repeated motif L=5, 400 times 'ATCGA'
MC0	markov chain of order 0, p_A=0.50, p_T=0.30, p_C=0.15, p_G=0.05
MC1	markov chain of order 1, p(T|A)=p(C|T)=p(G|C)=p(A|G)=0.91; otherwise p(*|*)=0.03
Es	experimental promoter regions of B.subtilis - see text

Table 5.1: Sequence DNA dataset used in this study. Description of DNA sequences generated to test Rényi quadratic entropy, all having length $N = 2000$ symbols. Otherwise noticed, "random" denotes a DNA sequence generated with the same probability of each symbol, i.e. $p_A = p_T = p_C = p_G = 0.25$. For inserted motifs, both the motif inserted and their lengths L are specified. For Markov Chain (MC) models the corresponding transition probabilities used to generate the sequence are given.

uses a random DNA sequence with insertions of specific exact motifs in known locations (*mo3*, *mo4* and *mo5*). For example in a sequence with length 2000, motifs were inserted in all the positions $50 + 100n$, $n = 0, \ldots, 19$. A motif of length 7 with a substitution or error in the middle position (*mo7e*) was also used. Another sequence category consisted in the repetition of the same symbol (*R1*) or motif (*R5*) the exact number of times to reach 2000. Additional, sequences were generated based in Markov chains of order 0 (*MC0*) and order 1 (*MC1*). Finally this study also includes real biological DNA sequences obtained from experimental data of promoter regions in B.subtilis (*Es*) (Helmann, 1995; Vanet et al., 1999). The tested sequence *Es* corresponds to the concatenation of 20 upstream regions before transcription, each with length 100, and all having a known promoter sequence constituted by the sub-string TTGACA—(space)—TATAAT with at most one substitution (known as the TATA-box). All the dataset and additional information are available in the webpage referred to above.

The tests consisted in calculating H_2 for different values of kernel variances σ^2 and analyze the curves obtained. The H_2 values were further compared with the quantiles of the empirical distributions of the random model.

5.3 Results and Discussion

This section presents the Rényi continuous quadratic entropy results H_2 obtained for the sequences dataset described above (see Systems and Methods) and further compares H_2 with a null model that corresponds to random sequences obtained by simulation. The relations between this new measure and the discrete Shannon's L-block entropy are also investigated.

5.3.1 Rényi continuous quadratic entropies H_2

The following Fig. 5.2 represents the graph of the function H_2 vs. $\ln\sigma^2$ for all the sequences in the dataset. As expected, the deterministic sequences have

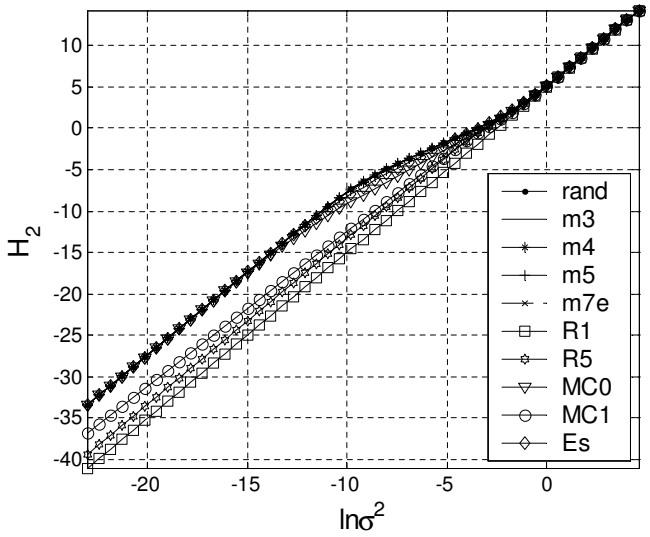


Figure 5.2: Rényi continuous quadratic entropy for the sequence DNA dataset. Representation of entropies for the dataset described in Tab. 5.1 as a function of the logarithm of the gaussian kernel variance $\ln\sigma^2$ used in the Parzen's Method. The lower the value of entropy H_2, the less random or more structured the sequence is. Graph has theoretically demonstrated asymptotes for $\ln\sigma^2 \to +\infty$ given by line $H_2^+ = 2\ln\sigma^2 + ln16\pi^2$ and for $\ln\sigma^2 \to -\infty$, line $H_2^- = 2\ln\sigma^2 + ln16\pi^2 + \ln N$ (see text).

lower entropy values, which corresponds to higher redundancy. The sequence *R1*, where the same symbol 'A' is repeated 2000 times, has the lower entropy measure for all kernel variances σ^2, thus representing a minimum threshold value for H_2. The maximum entropy values are attained for random sequences (*rand*), as would also be expected. The asymptotic behavior deduced in the Systems and Methods section is illustrated in this plot with all the sequences' Rényi entropies approaching the straight line given by Eq. 5.8 when $\ln\sigma^2 \to +\infty$.

For reference, the graph of H_2 for random sequences of different lengths N is represented in Fig. 5.3. These results were obtained by simulation, and the values represented correspond to the median entropies H_2 of all the replicates. Once more the asymptotic behavior of H_2 when $\ln\sigma^2 \to +\infty$ is confirmed, independently of the sequence length. Additionally, it is also apparent the linear asymptotic behavior for $\ln\sigma^2 \to -\infty$, for which there is an approximately linear relationship between H_2 and $\ln N$, as deduced: the longer the sequence, the higher its entropy is. This result provides the foundation to construct a simplified model of Rényi entropies as a function of the sequence length and kernel variance used.

In order to study how these values relate to one particular σ^2 an approximate derivative of the H_2 was obtained. This procedure is analogous to the calculation

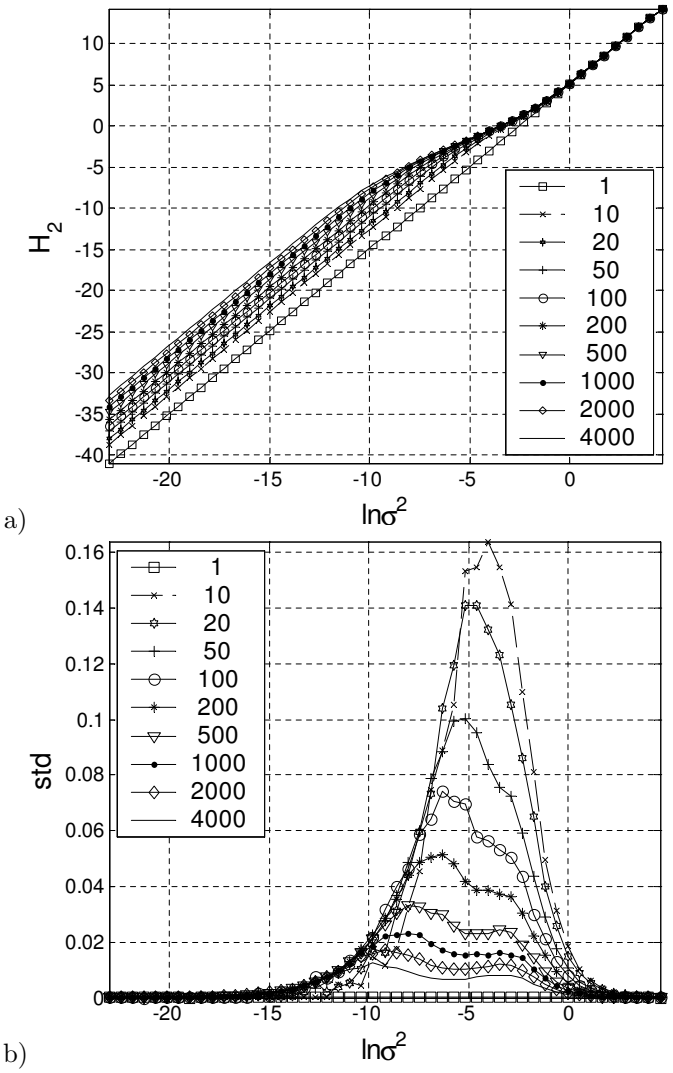


Figure 5.3: Rényi continuous quadratic entropy for random simulated sequences. a) This figure presents the Rényi continuous quadratic entropy H_2 for random sequences of different lengths: $N = 1, \ldots, 4000$ (see plot legend). The results correspond to the median H_2 values obtained in the simulation of 1000 replicates (random sequences) with specified length N. The longer the random sequence, the higher its entropy. Smaller sequences have fewer degrees of freedom, which is reflected in lower entropy values. The graph asymptotic behavior for $\ln \sigma^2 \to +\infty$ is the same as for the previous Fig. 5.2 and independent of N, as deduced (see text) – with line $H_2^+ = 2 \ln \sigma^2 + ln16\pi^2$. For $\ln \sigma^2 \to -\infty$ the asymptote is $H_2^- = 2 \ln \sigma^2 + ln16\pi^2 + \ln N$. b) The standard deviations of the H_2 values have a two-mode distribution for some N, and have limiting values 0 for $\ln \sigma^2 \to \pm\infty$.

of differential or conditional L-block Shannon entropies in the discrete case, given by $\Delta H^{(L)} = H^{(L+1)} - H^{(L)}$ (Schmitt and Herzel, 1997). There is an analogy between L-tuple value and kernel variance here explored. Figure 5.4 shows the first order differences between H_2 values, normalized by σ^2 steps. In this representation it is possible to distinguish more clearly the differences

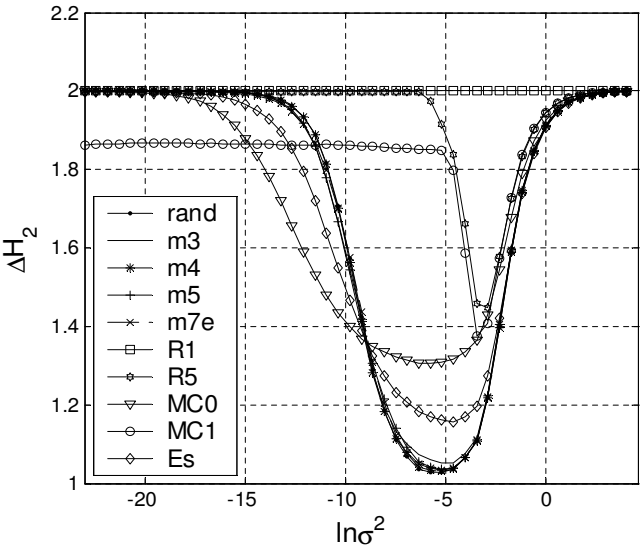


Figure 5.4: Derivative of Rényi quadratic entropy for the sequence DNA dataset. The values were obtained as the first derivatives of the Rényi entropy results represented in Fig. 5.2.

between the sequences. For example regarding the biological sequence *Es*, which in the previous Fig. 5.2 was superimposed with all the others, in Fig. 5.4 a clear difference is spotted for its first derivatives, suggesting a new aspect of entropy variation with potential use for discrimination. It is noteworthy that the maximum redundancy sequence *R1*, composed of a 2000 long sequence of 'A', has the same derivative as the asymptote, with value 2.

Analogously, Figure 5.5 contains the Rényi entropy derivatives obtained for the random sequences simulation, which confirms deviation from the asymptotic slope 2 as a measure of disorder. Figure 5.5 also illustrates the relation between the minimum value of the entropy derivative ΔH_2 and the length of the random sequences: longer sequences have lower minima, reached in a lower value of σ^2. This observation might represent the continuous counterpart to the discrete finite sample effect described in the introductory section. Furthermore this result can establish the possibility of measuring information content as the equivalent length of a random sequence, which will be explored further below.

In order to produce a quantitative representation of the comparison with the uniformly random reference, the corresponding quantile order of H_2 was calculated, as compared with the simulations performed for random sequences represented in Fig. 5.3. This approach can be interpreted as calculating the p-value of the test that each sequence is random. Figure 5.6 plots the quantile order values, always referred to the null random model, for each sequence

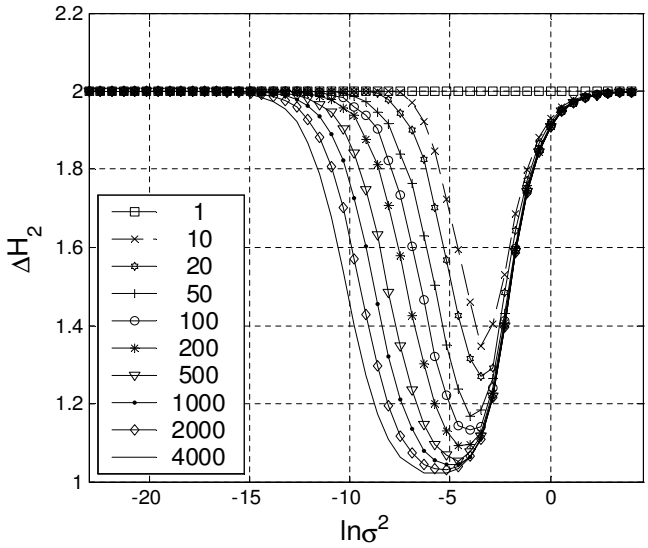


Figure 5.5: Derivative of Rényi quadratic entropy for random simulated sequences. Corresponds to the derivative of Rényi values obtained in the simulations (Fig. 5.3). The longer is the sequence the lower is the minimum value of the derivative, which corresponds to the slope of the entropy values $H_2(\ln \sigma^2)$.

vs. $\ln \sigma^2$. The random sequences have quantile order values near 0.5 (median). The sequences with motifs have values near 0 for most of the σ^2. The oscillations for values of $\ln \sigma^2 \simeq 10$ correspond to higher standard deviations of the simulated H_2 values, which have a two-mode distribution for some N (Fig. 5.3). One particularly interesting result observed in this figure was the quantile oscillation of the biological sequence *Es*: significantly, the promoter sequences are more deterministic than random for lower σ^2 values but for higher values of the kernel variance their H_2 value is above the maximum value obtained in the simulations (quantile probability near 1). These profiles offer a reference to calibrate observed sequence randomness at different resolution levels.

5.3.2 Equivalent sequence length N_{eq}

As noted above, the randomness of a sequence can be represented by an equivalent random sequence length N_{eq}, estimated by interpolation of the entropy values for the simulated random case (see Fig. 5.3a). This value corresponds to the length of the random sequence whose entropy is the same as the target sequence. For example, a sequence composed of only one symbol repeated, such as *R1*, has an entropy profile equal to the random sequence with just one symbol. Analogously sequence *R5* has the same entropy values than a random 5-symbol sequence, i.e. $N_{eq} = 5$, which corresponds precisely to the length of the repeated motif, showing that Rényi entropy H_2 has captured the redundancy of each sequence. For the other sequences in the dataset (see Tab. 5.1 for a description) the N_{eq} values depend on the kernel variance σ^2, which illustrates the fact that for all but the most redundant or random sequences, the equivalent

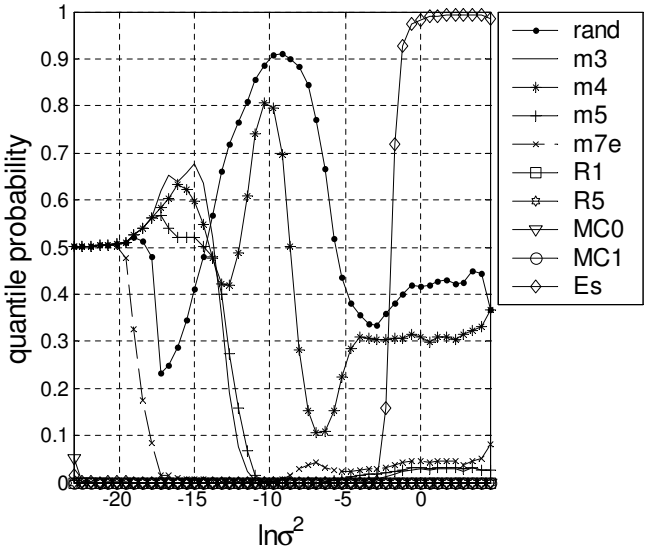


Figure 5.6: Quantile order or probability values of Rényi entropy for the sequence DNA dataset. Quantile order obtained when comparing each H_2 value with the distribution of 10^4 simulated random sequences of the same length $N = 2000$. Analogy to a p-value in the test for randomness. The values near zero correspond to sequences with entropy values below the minimum obtained for random sequences of the same length. The values near 0.5 are the median obtained in the simulation. Note the oscillating behavior of *Es*, with discrepant quantile values for different σ^2.

random length is contingent on the resolution considered.

5.3.3 Comparison between continuous and discrete measures of entropy

In this section the relationship between Rényi continuous entropy H_2 and the reference discrete Shannon's L-block entropy is investigated. The values for Shannon's entropy for the sequence dataset used are represented in Fig. 5.7. This plot shows clearly the finite sample effect: for infinite random sequences the entropy should follow a linear relationship (dashed line in the graph) but instead a decay of entropy values for L-tuple above 6 is observed. This can be quickly reasoned by noting that above this resolution there are actually more possible outcomes 4^6 than the possible observed frequencies of L-tuples. For higher L there is one point *per* tuple box, which corresponds to maximum entropy as a function of sequence length N, given by $H_{\max} = \ln N$.

In order to assess the association level between both measures, the correlation coefficient between discrete and continuous entropy values for the sequence dataset was calculated. (Using the quadratic discrete entropy, $\alpha = 2$, and not the common Shannon entropy, $\alpha = 1$, since the goal was to compare measures and not the effect of α on the entropy results.) The high correlation obtained for specific (L-tuple, σ^2) pairs shows that Rényi continuous entropy is linearly related to Shannon's L-block entropy. For example the correlation is maximal

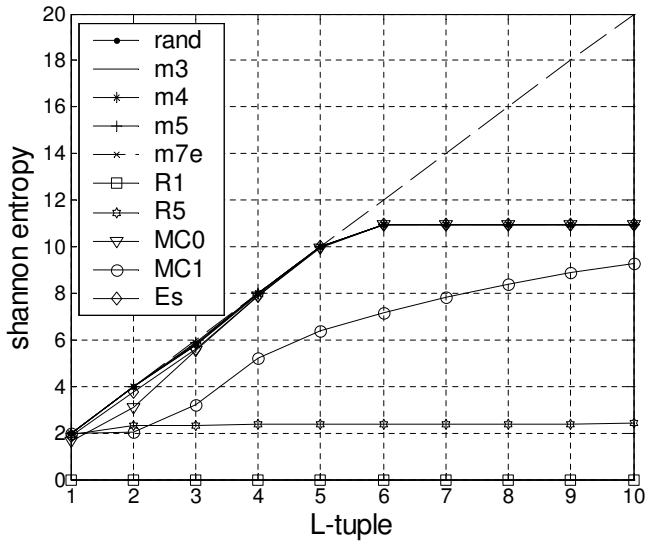


Figure 5.7: Rényi or Shannon discrete entropies for the sequence DNA dataset. Calculated with the formula analogous to Shannon's entropy, for each L-tuple frequencies. The levelling of entropy values for high values of L reflects the finite sample effect, which corresponds to the underestimations of the entropy for large L: random infinite-length sequences have entropies that match the dashed diagonal line (– –).

for the pair (6-tuple, 5.6e-8), for which a linear correspondence is given by the regression line $H_2 = 0.6925 H_{Shannon} - 28.3175$ with $R^2 = 1$ (graph not shown). This establishes that continuous measures of entropy are smoothly connected with Shannon's discrete values, of which Rényi entropies represent an extension to a more general solution.

5.3.4 Algorithm implementation – Rényi-Toolbox

Table 5.2 briefly describes the MATLAB functions made available online to perform all the calculations.

5.4 Conclusions

In this report a novel entropy measure of DNA sequences is proposed. This measure is based on the Rényi quadratic entropy definition for continuous probability functions and is used in conjunction with the Parzen window method, applied to the point density estimation of the Chaos Game Representation/Universal Sequence Maps (CGR/USM) of a sequence. It was verified that continuous Rényi quadratic entropy is a good measure of randomness of DNA sequences by testing the method both in artificial and biological sequences. Although a finite size sample problem might arise, similarly to the discrete equivalent one seeks to expand, the continuous entropy measure proposed enriches sequence randomness determination by not requiring a L-tuple count assignment, in effect searching the kernel variance space continuously. By freeing the calculation

File name	Brief description
readfasta.m	Reads sequences from FASTA format files to structured MATLAB variables
usm_make.m	Creates CGR/USM coordinates of a sequence
sig2.mat	Variances σ^2 of the Gaussian kernel tested
renyi2usm_fast.m	Calculates Rényi quadratic entropy of CGR/USM coordinates for several σ^2 and saves results in file
simul_renyi_usm2.m	Rényi entropies for random sequences Montecarlo simulation (calls randUSM.m)
randUSM.m	Generates USM coordinates of random sequence
usm_entropy.m	L-tuple discrete Shannon's entropy (Calls entropy_renyi.m)
entropy_renyi.m	Calculates Rényi discrete entropy of a vector (counts or probabilities)
example.m	script with full example of application
RenyiManual.pdf	Brief example on how to use the functions described – uses example.m output

Table 5.2: Rényi MATLAB toolbox function description

of this discrete restriction, a deeper insight on the randomness level of a sequence is achieved, by simultaneously probing variable orders. Moreover, the simplifications obtained with Parzen method allow for its straightforward and efficient computation. The proposed Rényi continuous quadratic entropy provides new tools for the study of motifs and to describe general repeatability in biological sequences. Additionally, this technique can eventually be applied to the development of compression tools for DNA data.

5.5 Acknowledgements

The authors thankfully acknowledge the financial support by grants SFRH/-BD/3134/2000 to S.V. and SAPIENS/34794/99 from Fundação para a Ciência e a Tecnologia (FCT) of the Portuguese Ministério da Ciência e do Ensino Superior. The authors also thank Daniel Paulino, of the Instituto Superior Técnico – Lisboa, for insightful suggestions during the preparation of the manuscript and Simone Preziati for carefully reading the manuscript and providing helpful comments.

5.6 Appendix

A. Gaussian or Normal distribution function definition

The Gaussian or Normal p-dimensional distribution with mean μ and covariance matrix Σ is given by the following Eq. A.1, where $x \in \mathbb{R}^p$ is a p-dimensional

random vector, x^T is the transpose vector of x and $|\Sigma|$ is the determinant of Σ:

$$g_p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{A.1}$$

When a random variable X, taking values in $\mathbb{R}^p$, has a probability density function (pdf) given by the former equation we say that $X \sim N_p(\mu,\Sigma)$.

B. Rényi quadratic entropy simplification

When using the Rényi quadratic entropy with Gaussian kernels, there is an important simplification given the property of the convolution of two Gaussians being also a Gaussian, i.e. $\int g(x-a_i;0,\Sigma_1)g(x-a_j;0,\Sigma_2)dx = g(a_i-a_j;0,\Sigma_1+\Sigma_2)$ (full proof on page 121). Hence the following demonstration is obtained (Eq. B.1), when applying to the USM representation of a N-length sequence, where I_p is the $(p\times p)$-dimension identity matrix.

$$\begin{aligned}
H_2(USM) &= -\ln\int \hat{f}^2(x)dx = -\ln\int\left(\frac{1}{N}\sum_{i=1}^{N} g_4(x-a_i;0,\sigma^2 I_p)\right)^2 dx \\
&= -\ln\int \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} g_4(x-a_i;0,\sigma^2 I_p)g_4(x-a_j;0,\sigma^2 I_p)dx \\
&= -\ln\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int g_4(x-a_i;0,\sigma^2 I_p)g_4(x-a_j;0,\sigma^2 I_p)dx \\
&= -\ln\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} g_4(a_i-a_j;0,2\sigma^2 I_p)
\end{aligned} \tag{B.1}$$

This expression further simplifies to Eq. B.2 since the kernel used was the spherical 4-dimensional Gaussian distribution $g_4(x;0,2\sigma^2 I_4) = \frac{1}{16\pi^2\sigma^4}\cdot\exp\left(-\frac{1}{4\sigma^2}x^T x\right)$:

$$\begin{aligned}
H_2(USM) &= -\ln\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{16\pi^2\sigma^4}\exp\left(-\frac{1}{4\sigma^2}d_{ij}\right) \\
&= -\ln\frac{1}{N^2 16\pi^2\sigma^4}\left[2\cdot\sum_{\substack{i<j\\ i,j=1}}^{N}\exp\left(-\frac{1}{4\sigma^2}d_{ij}\right)+N\right]
\end{aligned} \tag{B.2}$$

where $d_{ij} = (a_i-a_j)^T(a_i-a_j) = \sum_{k=1}^{4}\left(a_i^{(k)}-a_j^{(k)}\right)^2$ represents the squared Euclidean distance between sample USM points a_i and a_j and the last simplification occurs because all the pairwise distances $d_{ii}=0$ and $d_{ij}=d_{ji}$.

C. Asymptote calculation

The graph of function $H_2 = H_2(\ln\sigma^2)$ has an asymptote for $\ln\sigma^2 \to +\infty$ given by the straight line $H_2^+ = m\ln\sigma^2 + b$, where m and b are given by

the following Eq. C.1 and C.2 – using the notation simplification $\nu = \sigma^2$ and $d_{ij} = (a_i - a_j)^T(a_i - a_j)$:

$$\begin{aligned} m &= \lim_{\nu\to+\infty} \frac{H_2(\nu)}{\ln\nu} \\ &= \lim_{\nu\to+\infty} \frac{-\ln\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{16\pi^2\nu^2}\exp\left(-\frac{1}{4\nu}d_{ij}\right)}{\ln\nu} \\ &= \lim_{\nu\to+\infty} \frac{\ln 16\pi^2N^2 + 2\ln\nu - \ln\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(-\frac{1}{4\nu}d_{ij}\right)}{\ln\nu} \\ &= 2 - \lim_{\nu\to+\infty} \frac{\ln\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(-\frac{1}{4\nu}d_{ij}\right)}{\ln\nu} \\ &= 2 \end{aligned} \tag{C.1}$$

And the parameter b is:

$$\begin{aligned} b &= \lim_{\nu\to+\infty} (H_2(\nu) - m\ln\nu) \\ &= \lim_{\nu\to+\infty} \left[-\ln\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{16\pi^2\nu^2}\exp\left(-\frac{1}{4\nu}d_{ij}\right) - 2\ln\nu\right] \\ &= \lim_{\nu\to+\infty} \left[-\ln\frac{\nu^2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{16\pi^2\nu^2}\exp\left(-\frac{1}{4\nu}d_{ij}\right)\right] \\ &= \lim_{\nu\to+\infty} \left[-\ln\frac{1}{16\pi^2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(-\frac{1}{4\nu}d_{ij}\right)\right] \\ &= \ln 16\pi^2 + \lim_{\nu\to+\infty} \left[-\ln\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(-\frac{1}{4\nu}d_{ij}\right)\right] \\ &= \ln 16\pi^2 \end{aligned} \tag{C.2}$$

since $\lim_{\nu\to+\infty}\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(-\frac{1}{4\nu}d_{ij}\right) = N^2$.

The graph asymptote is hence $H_2^+ = 2\ln\sigma^2 + \ln 16\pi^2$.

Analogously, for $\ln\sigma^2 \to -\infty$, i.e. $\sigma^2 \to 0$, the asymptote is given by the straight line $H_2^- = m'\ln\sigma^2 + b'$, where m' and b' are given by the following Eq. C.3 and C.4, using the same simplifications:

$$\begin{aligned} m' &= \lim_{\nu\to 0} \frac{H_2(\nu)}{\ln\nu} \\ &= 2 - \lim_{\nu\to 0} \frac{\ln\left(2\cdot\sum_{\substack{i<j\\ i,j=1}}^{N}\exp\left(-\frac{1}{4\sigma^2}d_{ij}\right) + N\right)}{\ln\nu} \\ &= 2 \end{aligned} \tag{C.3}$$

And the value b' is:

$$\begin{aligned} b' &= \lim_{\nu\to 0}(H_2(\nu) - m' \ln \nu) \\ &= \ln 16\pi^2 + \lim_{\nu\to 0}\left[-\ln \frac{1}{N^2}\left(2 \cdot \sum_{\substack{i<j \\ i,j=1}}^{N} \exp\left(-\frac{1}{4\sigma^2} d_{ij}\right) + N \right)\right] \\ &= \ln 16\pi^2 + \ln N \end{aligned} \tag{C.4}$$

since $\lim_{\nu\to 0}\left(2 \cdot \sum_{\substack{i<j \\ i,j=1}}^{N} \exp\left(-\frac{1}{4\sigma^2} d_{ij}\right) + N\right) = N$.

And hence the asymptote for $\sigma^2 \to 0$ is $H_2^- = 2\ln\sigma^2 + \ln 16\pi^2 + \ln N$.

D. Convolution of normal distribution functions

Convolution definition

The convolution $f * g$ of two functions $f(x)$ and $g(x)$ defined in $\mathbb{R}$ is given by:

$$f * g(z) = \int_{\mathbb{R}} f(x)g(z-x)dx \tag{D.1}$$

Convolution properties

Some of the properties of $f * g$ are described below.

1. $f * g = g * f$ (commutative);
2. $f * (g * h) = (f * g) * h$ (associative);
3. $f * (g + h) = (f * g) + (f * h)$;
4. $\frac{d(f*g)}{dx} = \frac{df}{dx} * g = f * \frac{dg}{dx}$;
5. $\int f * g = \int f \cdot \int g$;
6. laplace transform[2] $\mathcal{L}[f * g] = \mathcal{L}(f)\mathcal{L}(g)$;
7. in probability theory, the convolution of two functions has a special relation with the distribution of the sum of two independent random variables. If the two random variables X and Y are independent, with pdf f and g respectively, the distribution $h(z)$ of $Z = X + Y$ is given by $h(z) = f * g$. This result is obtained below:

[2]Laplace transform of function $f(t)$ is defined as $\mathcal{L}[f(t)](s) = \int_0^\infty f(t)e^{-st}dt$

$$\begin{aligned} H(z) &= P(Z \leq z) = P(X+Y \leq z) \\ &= \int P(X+Y \leq z | Y = y) \cdot g(y) dy \\ &= \int P(X \leq z-y) \cdot g(y) dy \\ &= \int F_X(z-y) \cdot g(y) dy \end{aligned}$$

$$\begin{aligned} h(z) &= \frac{dH(z)}{dz} = \frac{d(\int F_X(z-y) \cdot g(y) dy)}{dz} \\ &= \int \frac{d(F_X(z-y))}{dz} \cdot g(y) dy \\ &= \int f(z-y) \cdot g(y) dy \\ &= f * g \end{aligned}$$

Convolution of normal distribution functions

Given two p-dimensional normal probability density functions represented as $G_1 \equiv g_p(x; a, A)$ and $G_2 \equiv g_p(x; b, B)$ (see Eq. A.1) we will prove that the convolution of these two functions is a normal probability density distribution function with mean $a+b$ and variance $A+B$, i.e. $g_p(x; a+b, A+B)$:

$$G_1 * G_2(z) = g_p(z; a+b, A+B)$$

The next sections demonstrate this result by first presenting an algebraic simplification of integrals using some properties of determinants and the factorization of quadratic forms.

Integral simplification

The following deduction represents the simplification of the integral $\int G_1 \cdot G_2 dx$ where G_1 and G_2 are the pdf of the normal distribution described above.

$$\begin{aligned} & \int g_p(x; a, A) \cdot g_p(x; b, B) dx \\ = & \int \frac{1}{(2\pi)^{p/2} |A|^{1/2}} e^{-\frac{1}{2}(x-a)' A^{-1} (x-a)} \frac{1}{(2\pi)^{p/2} |B|^{1/2}} e^{-\frac{1}{2}(x-b)' B^{-1} (x-b)} dx \\ = & \int \frac{1}{(2\pi)^{p/2} |A|^{1/2}} \frac{1}{(2\pi)^{p/2} |B|^{1/2}} e^{-\frac{1}{2}\left((x-a)' A^{-1} (x-a) + (x-b)' B^{-1} (x-b)\right)} dx \\ = & \int \frac{1}{(2\pi)^{p/2} |A|^{1/2}} \frac{1}{(2\pi)^{p/2} |B|^{1/2}} e^{-\frac{1}{2}\left((x-c)' \left(A^{-1}+B^{-1}\right)(x-c) + (a-b)' C (a-b)\right)} dx \\ = & \frac{\left|\left(A^{-1}+B^{-1}\right)^{-1}\right|^{1/2}}{(2\pi)^{p/2} |A|^{1/2} |B|^{1/2}} e^{-\frac{1}{2}(a-b)' C (a-b)} \cdot \\ & \qquad \cdot \int \frac{1}{(2\pi)^{p/2} \left|(A^{-1}+B^{-1})^{-1}\right|^{1/2}} e^{-\frac{1}{2}(x-c)' \left(A^{-1}+B^{-1}\right)(x-c)} dx \end{aligned}$$

$$
\begin{aligned}
&= \frac{\left|\left(A^{-1}+B^{-1}\right)^{-1}\right|^{1/2}}{(2\pi)^{p/2}\left|A\right|^{1/2}\left|B\right|^{1/2}} e^{-\frac{1}{2}(a-b)'C(a-b)} \qquad \text{(D.2)}\\
&= \frac{1}{(2\pi)^{p/2}\left(|A|\,|B|\,|A^{-1}+B^{-1}|\right)^{1/2}} e^{-\frac{1}{2}(a-b)'(A+B)^{-1}(a-b)}\\
&= \frac{1}{(2\pi)^{p/2}\left|ABA^{-1}+ABB^{-1}\right|^{1/2}} e^{-\frac{1}{2}(a-b)'(A+B)^{-1}(a-b)}\\
&= \frac{1}{(2\pi)^{p/2}\left|ABA^{-1}+A\right|^{1/2}} e^{-\frac{1}{2}(a-b)'(A+B)^{-1}(a-b)}\\
&= \frac{1}{(2\pi)^{p/2}\left|A\left(B+A\right)A^{-1}\right|^{1/2}} e^{-\frac{1}{2}(a-b)'(A+B)^{-1}(a-b)}\\
&= \frac{1}{(2\pi)^{p/2}\left|A+B\right|^{1/2}} e^{-\frac{1}{2}(a-b)'(A+B)^{-1}(a-b)}
\end{aligned}
$$

Properties of determinants

1. $|AB| = |A|\,|B|$
2. $\left|A^{-1}\right| = \frac{1}{|A|}$
3. $|cA| = c^n\,|A|$
4. $\left|BAB^{-1}\right| = |B|\,|A|\,\left|B^{-1}\right| = \frac{|B||A|}{|B|} = |A|$
5. $\left|B^{-1}AB-\lambda I\right| = \left|B^{-1}AB-B^{-1}\lambda IB\right| = \left|B^{-1}\left(A-\lambda I\right)B\right| = |A-\lambda I|$

Factorization of quadratic forms

Given x, a and b vectors of dimension p, A and B symmetric matrices of order p positively defined such as $A+B$ is not singular, we have the following result, demonstrated e.g. in (Box and Tiao, 1973):

$$
\begin{aligned}
(x-a)'\,A\,(x-a)+(x-b)'\,B\,(x-b) = \\
(x-c)'\,(A+B)\,(x-c)+(a-b)'\,C\,(a-b) \qquad \text{(D.3)}
\end{aligned}
$$

$$
\begin{aligned}
\text{where} \qquad c &= (A+B)^{-1}\,(Aa+Bb)\\
C &= A\,(A+B)^{-1}\,B = \left(A^{-1}+B^{-1}\right)^{-1}
\end{aligned}
$$

Result

Let $G_1(x)$ and $G_2(x)$ be the probability density function of the p-dimensional normal distributions $N(a,A)$ and $N(b,B)$ respectively. The convolution $G_1 * G_2$ is defined as:

$$
\begin{aligned}
G_1 * G_2(z) &= \int G_1(x) G_2(z-x) dx \\
&= \int \frac{1}{(2\pi)^{p/2} |A|^{1/2}} e^{-\frac{1}{2}(x-a)' A^{-1}(x-a)} \cdot \\
&\qquad \cdot \frac{1}{(2\pi)^{p/2} |B|^{1/2}} e^{-\frac{1}{2}(z-x-b)' B^{-1}(z-x-b)} dx \qquad \text{(D.4)} \\
&= \int g_p(x; a, A) \cdot g_p(x; z-b, B) dx \\
&= \frac{1}{(2\pi)^{p/2} |A+B|^{1/2}} e^{-\frac{1}{2}(z-(a+b))'(A+B)^{-1}(z-(a+b))} \\
&= g_p(z; a+b, A+B)
\end{aligned}
$$

This means that the convolution $G_1 * G_2(z)$ is the pdf of the normal distribution $N(a+b, A+B)$.

5.7 References

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., and Fletcher, M. (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics*, 17(5):429–437. 103, 105

Almeida, J. S. and Vinga, S. (2002). Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*, 3(1):6. 103, 104, 105

Ash, R. B. (1990). *Information Theory*. Dover Publications, New York. 106

Barnsley, M. F. (1998). *Fractals Everywhere*. Academic Press, Boston. 103, 104

Berthelsen, C. L., Glazier, J. A., and Skolnick, M. H. (1992). Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Phys Rev A*, 45(12):8902–8913. 103

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley-Interscience. 122

Chechetkin, V. and Lobzin, V. (1996). Levels of ordering in coding and non-coding regions of DNA sequences. *Physics Letters A*, 22:354–360. 102

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York. 106

Crochemore, M. and Verin, R. (1999). Zones of low entropy in genomic sequences. *Comput Chem*, 23(3-4):275–82. 103

Deschavanne, P., Giron, A., Vilain, J., Fagot, G., and Fertil, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*, 16(10):1391–1399. 103

Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A., and Ziv, J. (1995). On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proc 6th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'95)*, pages 48–57, San Francisco, CA. SIAM. 102

Gabrielian, A. and Bolshoy, A. (1999). Sequence complexity and DNA curvature. *Comput Chem*, 23(3-4):263–74. 103

Gatlin, L. L. (1972). *Information theory and the living system*. Columbia University Press, New York. 102

Gokcay, E. and Principe, J. C. (2000). A new clustering evaluation function using Renyi's information potential. In *IEEE Intl. Conf. on Acoustic, Speech and Signal Proc (ICASSP'00)*. 108

Gokcay, E. and Principe, J. C. (2002). Information theoretic clustering. *IEEE Transactions on Pattern analysis and machine intelligence*, 24(2):158–171. 108

Hariri, A., Weber, B., and Olmsted, J., r. (1990). On the validity of Shannon-information calculations for molecular biological sequences. *J Theor Biol*, 147(2):235–54. 103

Helmann, J. D. (1995). Compilation and analysis of *Bacillus subtilis* σ^{A}-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res*, 23(13):2351–60. 110

Herzel, H., Ebeling, W., and Schmitt, A. O. (1994a). Entropies of biosequences: The role of repeats. *Phys Rev E*, 50(6):5061–5071. 102

Herzel, H. and Grosse, I. (1995). Measuring correlations in symbolic sequences. *Physica A*, 216:518–542. 103

Herzel, H., Schmitt, A., and Ebeling, W. (1994b). Finite sample effects in sequence analysis. *Chaos, Solitons & Fractals*, 4(1):97–113. 102

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res*, 18(8):2163–2170. 103, 104

Jimenez-Montano, M. A., Ebeling, W., Pohl, T., and Rapp, P. E. (2002). Entropy and complexity of finite sequences as fluctuating quantities. *Biosystems*, 64(1-3):23–32. 102

Johnson, R. A. and Wichern, D. W. (1998). *Applied multivariate statistical analysis*. Prentice Hall, New Jersey, 4th edition.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. Dover Publications, New York. 102

Krishnamachari, A., Mandal, V. m., and Karmeshu (2004). Study of DNA binding sites using the Renyi parametric entropy measure. *J Theor Biol*, 227(3):429–436. 103

Lanctot, J. K., Li, M., and Yang, E.-h. (2000). Estimating DNA sequence entropy. In *Proc 11th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'00)*, pages 409–418, San Francisco, CA. SIAM. 102

Lio, P., Politi, A., Buiatti, M., and Ruffo, S. (1996). High statistics block entropy measures of DNA sequences. *J Theor Biol*, 180(2):151–60. 102

Loewenstern, D. and Yianilos, P. N. (1999). Significantly lower entropy estimates for natural DNA sequences. *J Comput Biol*, 6(1):125–42. 102

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., Simons, M., and Stanley, H. E. (1994). Linguistic features of noncoding DNA sequences. *Phys Rev Lett*, 73(23):3169–72. 102

Oliver, J., Bernaola-Galvan, P., Guerrero-Garcia, J., and Román-Roldán, R. (1993). Entropic profiles of DNA sequences through chaos-game-derived images. *J Theor Biol*, 160(4):457–70. 103

Orlov, Y., Filippov, V., Potapov, V., and Kolchanov, N. (2002). "Complexity": software tools for analysis of information measures of genetic texts. In *Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh, North Carolina, USA. 103

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076. 104, 107

Principe, J. C. and Xu, D. (1999). Information-theoretic learning using Renyi's quadratic entropy. In *1st Intl. Workshop on Independent Component Analysis (ICA '99)*, pages 407–412, Aussois, France. 108

Principe, J. C., Xu, D., and Fisher, J. W. I. (2000). Information-theoretic learning. In Haykin, S., editor, *Unsupervised adaptive filtering*, volume 1, pages 265–319. John Wiley & Sons, New York. 104, 108

Rényi, A. (1961). On measures of entropy and information. In *Proc. of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, volume 1, pages 547–561. University of California Press. 102, 103, 106

Rényi, A. (1966). Introduction a la théorie de l'information. In *Calcul des probabilités*. Dunod, Paris. 102, 103, 106

Sadovsky, M. G. (2003). The method to compare nucleotide sequences based on the minimum entropy principle. *Bull Math Biol*, 65(2):309–22. 103

Schmitt, A. O. and Herzel, H. (1997). Estimating the entropy of DNA sequences. *J Theor Biol*, 188(3):369–77. 102, 113

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656. 102, 106

Siegrist, K. (1997–2004). Virtual Laboratories in Probability and Statistics. Accessed: 28 Jul. 2004. <*http://www.math.uah.edu/stat*>.

Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., and Simons, M. (1999). Scaling features of noncoding DNA. *Physica A*, 273(1-2):1–18. 103

Strang, G. (1988). *Linear Algebra and Its Applications*. International Thomson Publishing, 3rd edition.

Tino, P. (1999). Spatial representation of symbolic sequences through iterative function systems. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 29(4):386–392. 103

Tino, P. (2002). Multifractal properties of Hao's geometric representations of DNA sequences. *Physica A*, 304(3-4):480–494. 103

Tino, P. and Dorffner, G. (2001). Predicting the future of discrete sequences from fractal representations of the past. *Machine Learning*, 45:187–217. 103

Troyanskaya, O. G., Arbell, O., Koren, Y., Landau, G. M., and Bolshoy, A. (2002). Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5):679–88. 103

Vanet, A., Marsan, L., and Sagot, M.-F. (1999). Promoter sequences and algorithmical methods for identifying them. *Res. Microbiol.*, 150:779–799. 110

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison–a review. *Bioinformatics*, 19(4):513–523. 103, 104

Vinga, S., Gouveia-Oliveira, R., and Almeida, J. S. (2004). Comparative evaluation of word composition distances for the recognition of SCOP relationships. *Bioinformatics*, 20(2):206–215. 104

Weiss, O., Jimenez-Montano, M. A., and Herzel, H. (2000). Information content of protein sequences. *J Theor Biol*, 206(3):379–86. 103

Xiao, M., Zhu, Z. Z., Liu, J., and Zhang, C. Y. (2002). A new method based on entropy theory for genomic sequence analysis. *Acta Biotheor*, 50(3):155–65. 102

Yu, Z. G., Anh, V., and Lau, K. S. (2004). Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *J Theor Biol*, 226(3):341–8. 103

Chapter 6

Final discussion

The previous chapters have described the development of biological sequence analysis techniques not dependent on alignment algorithms. The main applications of these alignment-free methods were presented and extensively reviewed. Additionally, a quantitative evaluation of the dissimilarity measures thus obtained was performed for the classification of proteins as specified by the SCOP database. This work has also extended chaos game representation (CGR) procedure, initially proposed only for DNA, to higher order alphabets, thus accommodating proteins, and finally applied this generalization, named Universal Sequence Maps (USM), to estimate DNA Rényi's continuous entropy.

An important contribution of the present thesis was the extensive revision of alignment-free techniques for sequence analysis and comparison, systematizing the methods currently published and providing a collection of bibliographic references. Emphasis was also put on nomenclature standardization, an essential step for the progression of the field. These applications have shown that mapping sequences onto n-dimensional vectors permits a flexible and elegant solution for sequence comparison and classification. Very recently, and following the publication of this work, some papers have explored these techniques and have also extended dissimilarity measures, explicitly expanding on our review (Pham and Zuegg, 2004; Liao et al., 2004; Zimmermann et al., 2004), which validate the initial proposal of unifying alignment-free techniques in one unique approach.

Other important achievement was the quantification of classification accuracy of the dissimilarity measures reviewed, defining a protocol that can be easily extended to accommodate new proposed measures. Additionally, this work introduced a novel metric, W-metric, that merges alignment-free distances and methods based on alignment by reusing concepts from linear algebra and evolutionary substitution matrices. Word composition-based metrics can provide efficient filters for sequence comparison, subsequently sending the corresponding results onto more specific alignment-based algorithms, thus improving computational speed. For example, new methods for multiple sequence alignment are based on this idea (Edgar, 2004a,b).

In the attempt of investigating alternative resolution-free representations, a special contribution of the work described in this thesis was the generaliza-

tion of CGR maps to higher-order alphabets through Universal Sequence Maps (USM), thus exploring a representation that has long been fruitful in time series prediction and Markov chains modelling.

Finally, another important achievement was the definition of a novel continuous uncertainty measure of DNA sequences that considers Rényi generalization of Shannon's formalism. The theoretical framework was established, with analytical deduction of asymptotical behavior, and simulation studies performed when a solution is not available in closed form. This constitutes the foundation work to future endeavors to investigate repeatability and uncertainty of biological sequences, adjusting the parameters of this estimation to additional features extraction. The study of randomness and order in DNA can reveal important properties of how genetic information is stored, transmitted and structured.

This work has shown that dissimilarity measures and vector representations of biological sequences are far from being fully explored. Other important aspect that emerged from the results included in this thesis was the modelling of the sequences themselves; the results obtained further ascertain that the succession order of biological sequences is not totally captured by Markov chain (MC) models and, most likely, real DNA and protein strings present a more complex arrangement which explains the wide functions they carry out in cells. The MC extension given by CGR/USM representation might be the solution for efficiently estimating order-free structure that is independent of specific L-tuple resolutions. Ideally, spanning different scales simultaneously would provide an insightful view of how biological sequences are organized and how they interact at different levels. Evidence suggests that the cell entirely exploits this structure, notoriously ambiguous, thus augmenting the scope of possible output molecules and functions, furthermore generating complex control systems. The redundancy present in several processes, from metabolic networks to protein synthesis, along with their apparent stochastic component, establish a dialectic between determinism and randomness, whose equilibrium is the key to fully explore DNA vast information potential, though using the same mechanisms and cell's machinery. Interestingly, this is precisely the notion conveyed by entropy measures, bridging molecular biology and information theory. The study of probabilistic systems entropy provides accurate measures of coding, transmission and information gain, with direct implication to biological systems and highly related to the study of sequences.

Future work envisaged will include extensions of dissimilarity measures based on vector operations in the alignment-free space. As seen, these vector representations of sequences have numerous advantages that fully justify their use for DNA and protein comparison, as well as for efficient implementation of several classification algorithms. Additionally, it is essential to probe structures beyond Markov models, in which manageable vector maps play a fundamental role. In particular, the CGR/USM provides an appealing representation of sequences useful for feature extraction via machine learning algorithms. For example, the direct application of artificial neural networks or support vector machines in this space should be investigated, as a method to predict crucial regions in biological sequences and their respective representation. Ideally, key molecular functions, such as promoters and transcription factors in DNA, would be recog-

nized through clustered positions in CGR/USM obtained with those algorithms.

One possible development in protein classification, that followed directly the study performed, is the optimization of quadratic forms for protein pre-filtering, searching the scoring matrix space for the best combination of values that leads to more accuracy in protein (family through class) relationship recognition. In fact, the W-metric here proposed was not proven to be the most discriminant. Due to the high dimensionality of data and parameters, the optimization of the matrix associated with the quadratic form might be better achieved with resource to artificial neural networks associated with genetic algorithms. As referred to before, the optimal quadratic form-based dissimilarity measure thus obtained can be used for efficiently pre-processing protein datasets and heuristic reasoning for the improvement of alignment-based algorithms.

Another topic that can be extended in the future is related to the continuous Rényi entropy proposed in the present work. In fact, only the basic introduction to this measure was presented, without a comprehensive description of its potential relevance for sequence analysis. For example, one obvious application of the Rényi continuous entropy is the definition of *entropic profiles*, thus describing local symbol-in-context information (in preparation). These entropic profile signatures, based on the kernel density estimation, provide local properties of the sequence, possibly related to the statistical significance of motifs and suffixes. The profile values are taken directly from the Parzen's estimation, hence depending on the kernel variance parameter. This characterization is important, in this context, as a method to reveal how strongly statistical properties may convey biological meaning. The link between entropy and significance is very important and should be further investigated. In fact, one epistemological key question in bioinformatics is how the computational analysis of sequences robustly coincides with biological reality, a question without a definite answer. Possible applications of this technique in the future include prediction problems, e.g. intron-exon recognition, and identification of long-range correlations in DNA. Moreover, the straightforward extension of this methodology to proteins should be analyzed, with possibly suitable use of compressed alphabets as to reduce space dimensionality of Rényi entropy calculations.

Another important issue that should be considered in the future is related to the kernel density estimation itself (of Parzen's method); the use of Gaussian functions was justified by computational reasons and accepted a priori without discussion, but other kernel functions should also be investigated. In fact, rectangular variable kernels might be more appropriate given USM intrinsic square geometry, instead of densities that clearly spread outside the original set. Furthermore, the parameter α might also be adjusted and investigated, analyzing how the resulting entropy is affected by its change, verifying if this value can be optimized in specific problems.

Another central topic will be the optimization and efficient implementation of the algorithms described, with its subsequent application to large datasets for whole genome analysis.

Bioinformatics and biological sequence analysis is evolving very fast and their future is, without any doubt, indissociable from the forthcoming discoveries of molecular biology. In the next years, it is foreseen that new emerging

paradigms will nourish bioinformatics exploits, and vice-versa, bioinformatics will also propose new theories to biologists, creating a knowledge loop with challenging problems posed in both directions. It is worth mentioning that nowadays it is believed that sequence *per se* is not sufficient to grasp the functioning of all molecular biology processes. The complex network referred to requires integrative techniques to bring together different levels of information, as illustrated by the emerging microarray analysis and proteomics fields.

Very recently, new discoveries have shown that several dogmas, thought to be universally true, have exceptions, casting some doubt on the comprehension of biological systems. For example, the recent interest in *epigenetics* – term defining all meiotically and mitotically heritable changes in gene expression that are not coded in the DNA sequence (Egger et al., 2004) – is bringing back the controversial (not to say heretical ...) issue of Lamarckian evolution. Apparently, processes occurring in an organism due to environmental factors, such as the activation or silencing of some genes, can be passed to the next generation, influencing their gene expression and consequently their phenotype (Pray, 2004). Additionally, the recent discovery of a new mechanism named RNA interference (RNAi) puts to doubt the central dogma itself (Henikoff, 2002), since small fragments of RNA can interfere with genes, creating a feedback system of gene expression control (Novina and Sharp, 2004). The function of junk DNA is still unknown but recently several hypotheses concerning its vital role were proposed (Pearson, 2004). Other authors suggest that ribosome does not represent a passive role, being a key molecule in all the cell's processes (Barbieri, 1985). All these aspects show the continuous progress of fundamental concepts, far from being totally established.

These facts suggest that in the future bioinformatics will develop in the direction of the integration of multi-object and relational information from different source types. The expected trend will be of the conjugation of new techniques based on the transcriptome —denoting the transcribed elements of genomes, which includes mRNA, proteome and regulatory network analysis, as increasingly framed by the field of systems biology. Although this research will incorporate diverse levels of information, biological sequence analysis will ultimately be the key, at a lower hierarchical level, for uncovering the complex control networks known to be present in all living cells and will also be a fundamental issue to understand the phenomena of the cell.

6.1 References

Barbieri, M. (1985). *The semantic theory of evolution*. Harwood Academic Publishers, Chur, Switzerland; New York. 130

Edgar, R. C. (2004a). Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Res*, 32(1):380–5. 127

Edgar, R. C. (2004b). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113. 127

Egger, G., Liang, G., Aparicio, A., and Jones, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–63. 130

Henikoff, S. (2002). Beyond the central dogma. *Bioinformatics*, 18(2):223–225. 130

Liao, B. Y., Chang, Y. J., Ho, J. M., and Hwang, M. J. (2004). The Uni-Marker (UM) method for synteny mapping of large genomes. *Bioinformatics*, 20(17):3156–65. 127

Novina, C. D. and Sharp, P. A. (2004). The RNAi revolution. *Nature*, 430(6996):161–164. 130

Pearson, H. (2004). 'Junk' DNA reveals vital role. *Nature*. 130

Pham, T. D. and Zuegg, J. (2004). A probabilistic measure for alignment-free sequence comparison. *Bioinformatics*, 20(18):3455–3461. 127

Pray, L. A. (2004). Epigenetics: Genome, meet your environment. *The Scientist*, 18(13):14–20. 130

Zimmermann, J., Lipták, Z., and Hazelhurst, S. (2004). A method for evaluating the quality of string dissimilarity measures and clustering algorithms for EST clustering. In *Proceedings of the Forth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04)*, pages 301–309, Taiwan. 127

Index

alignment
    multiple, 51
    pairwise, 12
aminoacid, 5, 6
asymptote, 118
AUC, 87

base
    A,C,G,T, 4
bioinformatics, 2
bit, 23
BLAST, 14
BLOSUM, 86

central dogma, 8, 130
CGR, **18**, 104
chaos game, 17
chromosome, 4
Clustal, 51
codon, 9
computational biology, 3
concave function, 26
convolution, 120

deletion, 11
dissimilarity
    angle, 44
    correlation, 42
    d2, 41
    Euclidean, 39
    evolutionary, 45
    Kolmogorov, 47
    Kullback-Leibler, 24, 38, 44
    Mahalanobis, 42
    standard Euclidean, 43
    USM, 46
    W-metric, 86, **86**, 89, 94
    weighted Euclidean, 40
distance, 15, 36, *see also* dissimilarity
DNA, 4
    junk, 4, 130
    replication, 8
dynamic programming, 12

EBI, 54
entropy, 23, 37
    conditional, 24
    joint, 24
    Rényi, 25, 26, 102, 103, **106**
    relative, 24
    Shannon, 23, 26, 37, 102, 106
enzyme, 5
epigenetic, 130
exon, 9

fractal, 16

gap, 12
gene, 4
genetic code, 9
genome, 4

hemoglobin, 7, 11, 18, **49**
homologous, 12

information theory, 22, 37, 102, 106
insertion, 11
intron, 9
iterated function system, 16

Kullback-Leibler, *see* dissimilarity

L-tuple, **36**, 37

Markov chain, 21
metric, 15
metric space, 15
mutation, 11
mutual information, 25

NASC, 47, 48, 54

Matlab functions, 48
webpage, 33

PAM, 86
Parzen's method, 107
protein, 49, 85
definition, 5
functions, 5
structure, 7, 85
synthesis, 8, 10
W-metric, 86

Rényi, *see* entropy
reading frame, 10
RNA, 4, 5
mRNA, 9
RNAi, 130
rRNA, 9
tRNA, 10
ROC curves, 87, **87**, 90

SCOP database, 28, 88, **88**, 91
scoring matrix, 86
Shannon, *see* entropy
Sierpinski triangle, 17
splicing, 9
stochastic process, 20
substitution matrix, 12

transcription, 9
transition probability, 21

USM, 68, **105**

vector map, 14

W-metric, *see* dissimilarity

Curriculum Vitae

Susana de Almeida Mendes Vinga Martins was born on the 28th July 1975 in Lisbon, Portugal. From 1993 to 1999 she frequented the Instituto Superior Técnico (IST), Universidade Técnica de Lisboa, where she obtained the degree in Mechanical Engineering – Automation and Robotics. During that period she attended Biomedical Engineering for one year within the scope of the ERASMUS Program, in the Politecnico di Milano, Italia and finished the Music/Piano Course in the Instituto Gregoriano de Lisboa. From 1996-98 and 1999-2000 she was a Mathematical Analysis III teaching assistant in IST. In the period 2000-2002 she frequented graduate courses in Probability and Statistics in the Mathematics Department (IST) finishing a Post-Graduation in that area.

In 2001 she started her PhD project on *Biological Sequence Analysis* under the supervision of Prof. Jonas Almeida in the Biomathematics Group of the Instituto de Tecnologia Química e Biológica/Universidade Nova de Lisboa (ITQB/UNL). During the period from 2001 to 2004 she visited the Biometry and Epidemiology Department of the Medical University of South Carolina (MUSC) in Charleston (USA). She also visited the Department of Biométrie et Biologie Évolutive in Université Claude Bernard, Lyon (France) where she worked under the supervision of Prof. Marie-France Sagot.

During her PhD project she presented posters in the following conferences: 9th International Conference on Intelligent Systems for Molecular Biology - (ISMB'2001) in Copenhagen, Denmark; 5th European Conference on Mathematical and Theoretical Biology (ECMTB'2002) in Milan, Italy; 2nd European Conference on Computational Biology (ECCB'2003) in Paris, France and in the joining conference (ISMB/ECCB'2004) in Glasgow, UK.

Vinga, S. and Almeida, J. S. (2004). Rényi continuous entropy of DNA sequences. *J Theor Biol*, 231(3):377–388.

Vinga, S., Gouveia-Oliveira, R., and Almeida, J. S. (2004). Comparative evaluation of word composition distances for the recognition of SCOP relationships. *Bioinformatics*, 20(2):206–215.

Gomes, A. R., Vinga, S., Zavolan, M. and de Lencastre H. (2005). Analysis of the genetic variability of virulence-related loci in epidemic clones of methicillin-resistant Staphylococcus aureus. *Antimicrob Agents Chemother*, 49(1):366–79.

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison–a review. *Bioinformatics*, 19(4):513–523.

Almeida, J. S. and Vinga, S. (2002). Universal sequence map (USM) of arbitrary discrete sequences. *BMC Bioinformatics*, 3(1):6.