

Family Specific Rates of Protein Evolution

Hans-Günther Luz

Februar 2005

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich für Mathematik und Informatik
der Freien Universität Berlin
vorgelegte
Dissertation

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Hans-Peter Herzel

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Hans-Peter Herzel

Datum der Disputation: 20. Dezember 2006

“Kollegen, wir sind ganz nahe dran, das Rätsel zu lösen”, ignorierte Justus die Einwände des Dritten Detektivs. “Und das wird auch höchste Zeit, denn wir wissen immer noch nicht, wo Morton steckt und ob er vielleicht in Gefahr ist. Heute Abend werden wir uns ein weiteres Puzzleteil holen.”

Bob brachte beim Abendessen kaum einen Bissen herunter. Seine Eltern waren hocherfreut, dass er endlich mal wieder nach Hause kam, doch er enttäuschte sie durch seine Schweigsamkeit und die ständigen Blicke auf die Uhr.

Aus: Die drei ???, Tödliche Spur

Contents

1	Introduction	7
1.1	Motivation	7
1.2	Overview	7
1.3	Acknowledgements	8
2	Preliminaries	9
2.1	Biological Background	9
2.1.1	Cells	9
2.1.2	Proteins	9
2.1.3	DNA and gene transcription	13
2.1.4	Protein synthesis and secretion	14
2.2	Molecular Evolution	16
2.2.1	Molecular Evolution	16
2.2.2	Genome evolution	17
2.2.3	Point Mutations	17
2.2.4	Amino Acid Replacement	17
2.2.5	Evolution of protein function	18
2.3	Computational Molecular Biology	20
2.3.1	Pairwise sequence alignment	20
2.3.2	Fast heuristic alignment, BLAST and FASTA	21
2.3.3	Profiles, PSI-BLAST	21
2.3.4	Hidden Markov Models	22
2.3.5	Clustering homologous sequences	24
2.3.6	Multiple sequence alignment	27
2.4	Molecular Phylogenetics	29
2.4.1	Inferring molecular phylogenies	29
2.4.2	The molecular clock, ultrametric and additive trees	29
2.4.3	Character- and distance based methods	30
2.4.4	Non-parametric bootstrapping	30
2.4.5	Markovian modeling of sequence evolution	31

3	Family specific rates of protein evolution	41
3.1	Rates of protein evolution	41
3.2	The data, orthologous families and alignments	42
3.3	The substitution model	44
3.4	Comparing pairwise evolutionary distances	44
3.5	Estimating family specific evolutionary rates	47
3.5.1	The tree length ratio	47
3.5.2	The Family Specific Rate (FSR)	48
3.5.3	The Likelihood Ratio	48
3.6	Results	51
3.6.1	FSR versus tree length	51
3.6.2	Rate distribution	51
3.6.3	Family Specific Rates and nonsynonymous nucleotide substitutions	53
3.6.4	Rates of essential genes, RNA interference	57
3.6.5	Protein interaction and the rate of evolution	58
3.6.6	Does protein function constrain the rate?	59
4	Protein evolution, the younger the faster	63
4.1	Metazoan radiation, the younger the faster?	63
4.2	Extra-cellular proteins	64
4.2.1	Inferring extra-cellular localization	64
4.2.2	Modern extra-cellular proteins are fast evolving	65
4.2.3	Evolution of protein tyrosine kinases (PTKs) and tyrosine kinase receptors (rPTKs)	65
4.3	Inferring the ancient origin of eukaryotic proteins	69
4.4	Taxon-specific rate distributions	72
4.4.1	Proteins evolved from primordial domains	72
4.4.2	The least common taxon of a SYSTERS cluster	73
4.4.3	The least common taxon according to domain architecture . . .	74
4.4.4	Taxon-specific rate distributions	74
4.4.5	Metazoan radiation and the invention of extra-cellular proteins .	75
4.4.6	The younger the faster?	77
4.5	Multigene families	79
4.5.1	Duplicated genes are more conserved	79
4.5.2	“Young” and “old” multigene families	79
4.5.3	Dating duplication events	82
5	Phylogenetic trees from multiple genes	89
5.1	Inferring genome phylogenies	89
5.2	The Path Length Ratio (PLR) method	90
5.3	The weighted Individual Protein (wIP) method	90
5.4	Results	91

6 Summary and conclusions	95
Bibliography	98
A Scatter plots	111
B Domain names	113
C Complete prokaryotic proteomes	115
D Erklärung zur Urheberschaft	117

Abbreviations and notations

BYr	Billions of years
<i>C</i>	<i>Caenorhabditis elegans</i>
<i>D</i>	<i>Drosophila melanogaster</i>
DNA	Deoxyribonucleic acid
EMP	Evolutionary Markov Process
<i>F</i>	<i>Fugu rubripes</i>
FSR	Family Specific Rate
<i>H</i>	<i>Homo sapiens</i>
IP	Individual Protein
LCT	Least common taxon
LRT	Likelihood ratio test
NJ	Neighbor-Joining
ML	Maximum likelihood
mRNA	Messenger RNA
PAM	Point accepted mutation, percent accepted mutations
PNJ	Profile Neighbor-Joining
PLR	Path Length Ratio
PTK	Protein Tyrosine Kinase
RNA	Ribonucleic acid
RNAi	RNA interference
rPTK	Receptor protein tyrosine kinase

$\hat{\lambda}_i$	Family Specific Rate
\hat{l}_i	Tree Length Ratio
t_j	Evolutionary distances, edge lengths
τ_j	Historical times
P	Probability transition matrix
π	Stationary distribution
Pr	Probability
\mathcal{L}	Likelihood
Q	Rate matrix
T	Tree topology
d_N	Number of nonsynonymous substitutions
d_S	Number of synonymous substitutions

1 Introduction

1.1 Motivation

The advent of molecular sequence data affects the theory of evolution like a continuing revolution. Already in the 1960ies Kimura demonstrated having just a few amino acid sequences available that the rate of protein evolution is much too high to be solely explained by the logic of natural selection. According to Kimuras theory of neutral evolution the majority of molecular changes in evolution are randomly fixed.

Current large scale nucleic acid sequencing projects abundantly produce sequence data. Analyzing the data is the primary challenge and a prerequisite for a better understanding of the forces that drive evolutionary processes.

1.2 Overview

In this thesis the rates of protein evolution are estimated and analyzed on a genomic scale for diverse eukaryotic model organisms. At first, Chapter 2 briefly reviews some basics of molecular biology and molecular evolution with a focus on topics that are relevant for the thesis. Methods, tools and information about certain database releases that were used and are referred to in the thesis are described in Section 2.3. Markov models of sequence evolution play a central role in the thesis. The basics of Markovian models that are used to model sequence evolution and to estimate evolutionary rates are reviewed in Section 2.4.5.

Chapter 3 empirically documents that rate variations among proteins are mainly driven by family specific effects. The measure of a family specific rate is meant to describe the slow or fast evolution of a gene family and averages over lineage specific rate variations. Having derived an overall rate distribution, the prevalent assumption that essential genes are more evolutionarily conserved than nonessential ones is put forward and supported.

The observation that modern extra-cellular proteins are subject to a rapid accumulation of mutations suggests the following hypothesis: Selective constraints act weaker and evolutionary rates are larger the more recent a protein emerged in evolution.

Chapter 4 describes the experiments that were performed to investigate the hypothesis.

In Chapter 5 two methods to infer an organismal phylogenetic tree are proposed and applied. Both methods account for rate variations among gene families.

1.3 Acknowledgements

I feel happy to have met and to have worked with my colleagues at the department of Computational Molecular Biology.

I am deeply grateful to Martin Vingron who constantly supported me. He introduced me to sequence analysis and gave me the opportunity to work in this field. The discussions with him form the basis of the thesis. I appreciate his guidance a lot.

Tobias Müller spent that much time and energy to ask, to listen, to answer and to explain. It was a pleasure to work with him.

Sincere thanks to Antje Krause. She was always helpful and I benefited from her experience and her assistance in sequence clustering.

Anja von Heydebreck provided valuable and practical hints. I enjoyed to discuss about tree reconstructions with her.

I thank Eike Staub very much. The conversations with him motivated and pushed biological questions.

I would like to thank Sven Rahmann for a lot of valuable discussions, for his frankness and his willingness to help.

I'd like to express my gratitude to our system administrators, in particular to Wilhelm Rüsing and to Peter Marquardt. They are miraculous.

I am very thankful to Jörg Schultz for his support in calculating SMART E-values, to Birgit Pils for her assistance in inspecting multiple alignments, to Thomas Manke for discussing interaction data sets, to Thomas Meinel for his help in using the SYSTERS taxonomy and to Heiko Schmidt for unraveling PUZZLE code.

Prof. Herzog is willing to write a referee report for the thesis, for which I am grateful.

Thanks a lot for proofreading the manuscript, Tobias Müller, Antje Krause, Sven Rahmann, Steffen Grossmann, Stefan Röpcke and Utz J. Pape.