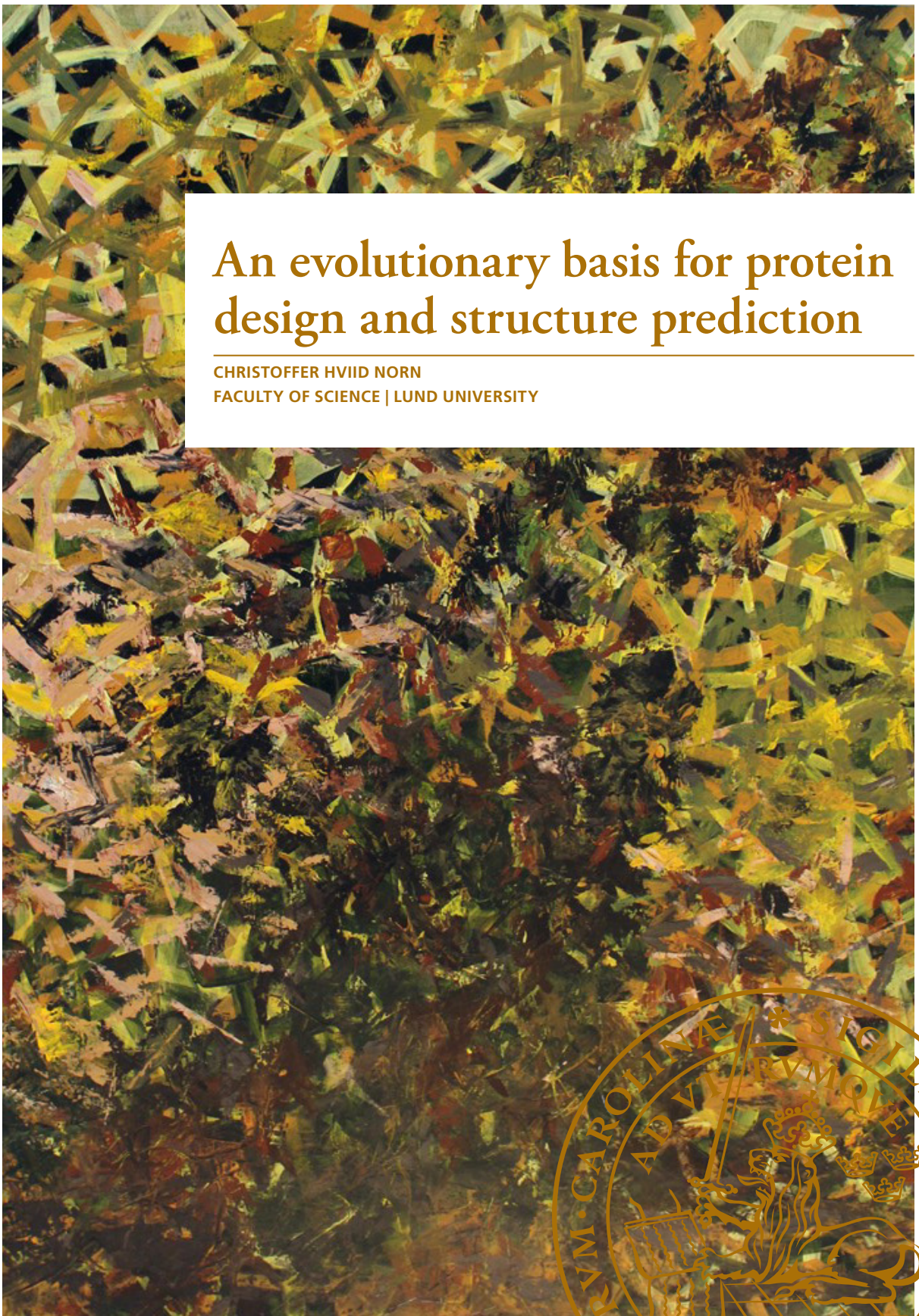
An evolutionary basis for protein design and structure prediction

Norn, Christoffer

2019



An evolutionary basis for protein design and structure prediction

CHRISTOFFER HVIID NORN
FACULTY OF SCIENCE | LUND UNIVERSITY

An evolutionary basis for protein design and structure prediction

An evolutionary basis for protein design and structure prediction

by Christoffer Hviid Norn

LUND
UNIVERSITY

Thesis for the degree of Doctor of Philosophy
Faculty opponent: Prof. David D. Pollock

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism in lecture hall B, Kemicentrum, Lund University on Friday, the 8th of February 2019 at 13:00.

Organization
LUND UNIVERSITY

Department of Biochemistry and Structural Biology
PO Box 124
SE–221 00 LUND
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2019-02-08

Sponsoring organization

Author(s)
Christoffer Hviid Norn

Title and subtitle
An evolutionary basis for protein design and structure prediction:

Abstract
The sequence diversity of protein families is a result of the biophysical selection pressures that shaped their evolutionary history. Among the dominant pressures is selection for protein thermostability, which in itself is an attractive target in protein engineering because of its importance for various biopharmaceutical properties, the performance of industrial enzymes, and the ability to design new protein functions.

In the first part of this thesis, we use models of evolutionary dynamics and biophysical fitness functions to derive the relationship between amino acid frequencies in sites of proteins and the stability effects of mutations. This analysis suggests that a commonly applied assumption (that amino acids frequencies are Boltzmann distributed) is inaccurate, and we provide a new relation consistent with the current understanding of evolutionary dynamics and protein fitness. Next, we study the extent to which the evolutionary pattern of amino acid substitutions can be explained by protein stability, as predicted using all-atom models of protein energetics. We show that at least 65% of the substitution pattern can be explained by thermostability. With the same model, we show that functional sites (e.g. active sites or binding sites) can be predicted when the apparent evolutionary site-rate deviates significantly from that of a stability-only null-model of evolution. Finally, we study how the strength of selective pressure affects the evolutionary behavior of proteins, again using the same models, but this time generating evolutionary trajectories. We find that energetic coupling between amino acids (coevolution) and the detriment of mutation increases as the strength of selection increases.

Antibodies are a key molecular component of the adaptive immune system of vertebrates and an important biopharmaceutical molecule. In the second part of the thesis, we predict and design the structure of antibodies by using energetics derived from sequence alignments and following the evolutionary encoded modular segmentation of the molecule. Through multiple design and test iterations, we were able to design antibodies, which express stably and, in some cases, bind target antigens. The developed structure prediction algorithm performs as well as other methods, is in some cases more accurate, and produces models with lower chemical strain. We use the structure prediction method to study a tumor-associated carbohydrate binding antibody.

Finally, we also review the literature on design of symmetrical protein self-assembly, and study the dynamical properties of a partially disordered chaperone protein, calreticulin.

Key words
Protein evolution, biophysics, protein design, protein structure prediction

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

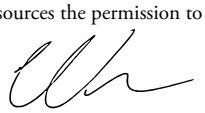
ISBN
978-91-7422-618-8 (print)
978-91-7422-619-5 (pdf)

Recipient's notes

Number of pages
247

Price

Security classification

Signature  Date 2019-01-02

An evolutionary basis for protein design and structure prediction

by Christoffer Hviid Norn

LUND
UNIVERSITY

Cover illustration front: Elizabeth Norn. Listen to the Light. 2016. Acrylic on canvas. Private collection.

Cover illustration back: Elizabeth Norn. Ambulo. 2016. Acrylic on canvas. Private collection.

Funding information: The thesis work funded in part by the Boehringer Ingelheim Fonds and the Swedish Research Council (2015-04203).

Faculty of Science, Department of Biochemistry and Structural Biology

ISBN: 978-91-7422-618-8 (print)
ISBN: 978-91-7422-619-5 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2019

To Tania, Viggo and Otto

Contents

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **A biophysical model of protein evolution relates amino acid frequencies and protein stability**
Norn, C., Theobald, D.L, and Andre, I.
Manuscript

II **A thermodynamic model of protein structure evolution explains amino acid rate matrices and predicts functional sites**
Norn, C., Theobald, D.L, and Andre, I.
Manuscript

III **Properties of all-atom simulations of protein evolution**
Norn, C., Theobald, D.L, and Andre, I.
Manuscript

IV **Computational design of protein self-assembly**
Norn, C. and Andre, I. (2016)
Curr. Opin. Struct. Biol. 39, 39–45

V **High-accuracy modeling of antibody structures by a search for minimum-energy recombination of backbone fragments**
Norn, C.H., Lapidoth, G., and Fleishman, S.J. (2017)
Proteins Struct. Funct. Bioinformatics, 85, 30–38

VI **Principles for computational design of binding antibodies**
Baran, D., Pszolla, M.G., Lapidoth, G.D., **Norn, C.**, Dym, O., Unger, T., Albeck, S., Tyka, M.D., and Fleishman, S.J. (2017).
Proc. Natl. Acad. Sci. 114, 10900–10905

VII **AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences**
Lapidoth, G.D., Baran, D., Pszolla, G.M., **Norn, C.**, Alon, A., Tyka, M.D., and Fleishman, S.J. (2015)
Sci. Rep. 8, 1–12

VIII **A combined computational-experimental approach to define the structural origin of antibody recognition of sialyl-Tn, a tumor-associated carbohydrate antigen**
Amon, R., Grant, O.C., Leviatan Ben-Arye, S., Makeneni, S., Nivedha, A.K., Marshanski, T., **Norn, C.**, Yu, H., Glushka, J.N., Fleishman, S.J., Chen, X., Woods, R.J., Padler-Karavani, V. (2018)
Proteins 83, 1385–1406

IX **Mapping the Ca2+ induced structural change in calreticulin**
Boelt, S.G., **Norn, C.**, Rasmussen, M.I., André, I., Čiplys, E., Slibinskas, R., Houen, G., and Højrup, P. (2016)
J. Proteomics 142, 138–148

Publications not included in this thesis:

Method of computational protein design
Fleishman, S., Lapidoth, G., Pszolla, M.G., **Norn, C.H**
Patent: WO2016005969A1

Acknowledgements

If you are new to science or have forgotten the feeling of having set sails towards a scientific destination only to get lost in a fog of uncertainty, I can tell you it is a daunting experience insurmountable without great company and good advice.

My Ph.D. started with my supervisor Ingemar, who in my pursuit of evolutionary-based protein-design of antibodies, helped me join an Israeli vessel already targeting that destination. Sarel, you taught me the beautiful mechanics of the boat-engine (Rosetta), what made it run, how to make it go through waters usually intractable, and the principles of structure. Thank you for the many thoughtful perspectives and the walks on the deck! I also thank Assaf, Shira, Ravit, and Adi who have been some of the most welcoming people I've encountered.

Before the destination was reached I changed boats. Returning to Sweden and Ingemar, I found myself on a different boat, which although equipped with a similar engine, also appeared to have a mast and a sail. To begin, we thought we had a well-defined destination, but it kept eluding us until it disappeared entirely. Another destination was set, but also that was elusive. At some point, lost in the fog, we started to study the sail and the wind that blew it. Without prior knowledge of wind dynamics, we were challenged. I thank you, Ingemar, for the many hours of helpful discussion, your patience, for being curious with me, trusting that studying the wind itself could be a destination on its own, and for your kindness. Also very important, thank you for agreeing that skånerost had no place on our boat and that we deserved better.

I thank you, Douglas, for providing clarity in foggy weather, your hospitality, and always checking that the foundations personal and otherwise were good. Anders, I thank you for introducing me to phylogeny and ancestral sequence reconstruction. Kresten and Amelie, I thank you both, for your continued interested in my doings, for always being up for discussions, and for your apparent encyclopedic knowledge of all things important. Charlotte for reading the entire thesis and discussing the projects as they both de- and evolved.

To the always kind people at the André lab. Wojtek especially for being of assistance scientifically and emotionally. Signe, I'm so sorry about the coffee! Mads for fun with metal-binding peptides. Filip for the well-ordered group meetings, Ryan for comforting me that politicians are not forever, and Camille and Caroline for making the group meetings tastier.

I'm also deeply grateful to the Boehringer Ingelheim Fonds and the people who run it. They supported part of my Ph.D. with a scholarship, multiple conferences, courses, and gave excellent carrier advice on the way!

Akos, Aslak, Charlotte, Damian, Harish, Jean, Jesper, Tabitha, and Tobias; my always supportive family; my dear Tania, Viggo, and Otto. This had, needless to say, not been possible without you.

Popular summary in English

Some 4 billion years ago, nature began the experiment that led to you, that bird on the branch, that branch, and everything else which is alive that you see around you. The generating process of all this splendid life is, of course, evolution. Evolution works not only on the macroscopic scale (think eyes, muscles, wings) but also on the molecular scale (think molecular antennas, nanoscale muscles, nanoscale motors). On the molecular scale, almost everything with a function is made of proteins and by proteins. Understanding how proteins work is, therefore, a major goal in science. They also play a crucial role in society: Our ability to engineer and design them is a key reason why many cancers are no longer a death sentence, why we can cold-wash your laundry (yearly saving the environment from millions of tonnes of CO_2), and it provides essential research tools for understanding the mechanics of life. However, proteins are not easily engineered. They do not have the same elegance as the famous DNA double-helix. Instead, there are, in the words of one of the discoverers of the hemoglobin structure (the protein that transports oxygen throughout your body), "hideous and visceral-looking"[1] objects.

To improve our ability to design proteins, we looked to nature for advice. Imagine visiting the butterfly collection at the Geological Museum at the University of Copenhagen, with the goal of altering the wing of a butterfly to improve its flight capabilities. You are met with a menagerie of thousands of wing shapes and body sizes. Which shape gives the fastest flight? Do you just "average" the wing-shape or do you take the most common shape? Neither. You need to know how flight capability affects survival and a model of how survival affects the observed butterfly diversity. In this thesis, we pursued a similar path, but for proteins. They are not kept at display in a traditional museums, but are stored in digital libraries accessible to anyone with an internet connection. We found that most of the variation in proteins could be explained by their stability (in the butterfly analogy, that flight capability is a major determinant of survival). Stability is an essential property for that protein engineers seek to optimize. We further found that stability could be predicted from the observed diversity.

Using the above knowledge and structure based models of protein stability, we designed a type of protein called antibodies. Antibodies are the reason that your body, most of the time, can defend itself against the festering of bacteria, virus, fungi, and cancers. They are also the reason that the biopharmaceutical industry earns some 200 billion dollars each year. We designed new antibodies that could bind two protein targets, and developed a new method, which could predict the structure of antibodies.

[1] Max Perutz

Chapter 1

Introduction

Nothing in Biology Makes Sense Except in the Light of Evolution[1]

Some 4 billion years ago life emerged on Earth [2] marking the beginning of one of the greatest natural experiments of all — evolution. Proteins soon became the main component of the molecular machinery of life and remained so ever since. Aside from being essential to all life forms, proteins were also co-opted by us humans for medical and technological purposes: We use them in vaccines to evolve our immune system to resist pathogenic invasions, in biopharmaceuticals to target cancers, and in industry to make more environmentally friendly processes. Over the past century, we interrogated the expanse of natural protein diversity, and we know today the sequence, structure, and function of millions of proteins, all connected through the process of evolution that built the tree of life. How can we use this information to understand evolution and improve or enable new medical and technological protein applications?

In this thesis, we studied the origin of patterns in protein evolution (paper I). Encouraged by the apparent simplicity we developed a new method to simulate protein evolution (paper II). We asked how one best interprets natural protein diversity to improve protein thermostability (paper III) — an essential parameter both in natural evolution and in industry. We developed a new evolutionarily-consistent method to predict the structure of antibodies, a type of immune system proteins (paper IV). We applied the prediction method to understand the molecular details of an antibody with relevance for cancer therapeutics (paper V). Using the evolutionary lessons we had yet to fully understand (paper III), we used the natural sequence and structure diversity of antibodies to develop an antibody design algorithm (paper VI) that proved successful in creating new protein-binding antibodies (paper VII).

[1] Christian Theodosius Dobzhansky [1].

Chapter 2

Evolution and evolutionary dynamics

All models are wrong, but some are useful [1]

Since Charles Darwin in 1859 described his theory of biological evolution [5] (also see [6]), which explains how species arise through natural selection between varying individuals with inheritable traits, much progress has been made. In the 1920s and 1930s the theory was married to Mendelian genetics by Roland Fisher, J. B. S. Haldane, and Sewall Wright and further expanded by Motoo Kimura in the 1950s, who realized that most observed mutations are the result of random genetic drift [7, 8]. In 1953, with the structure of DNA by Rosalind Franklin, James Watson, and Francis Crick, the physical basis of the theory was established [9]. Evolution has morphed from being a qualitative theory, to a quantitative theory, which is explained in the language of biology, mathematics and biophysics.

This chapter touches on the foundations of the theory applied in paper I-III. We discuss the origin and consequence of mutations, the fate of mutations, and the distribution of diversity.

2.1 The origin of mutations and effect of mutation

The three basic building blocks of evolution are mutation, replication and selection. Mating (recombination) does not result in new (or, as Darwin had struggled, loss of) genetic variation [10, 11]. Instead, mutations are the ultimate source of innovation on which evolution can act. Mutations occur in genetic material due to intrinsic

[1] [3, 4] George Box. In constructing models of evolution it is clear that, the complete "true" model is impossibly complicated, and would be as helpful as a 1:1 map of navigating the London Underground (that is, not at all). The question whether useful models that capture the essential patterns of evolution can be constructed?

instability of the nucleotides, environmental factors including radiation damage and chemicals, and the finite fidelity of the replication machinery. Genetically, there can be several consequences of a mutation event: A large proportion of mutations are point mutations, where a single base-pair is altered in the DNA [12], but mutations can also be tandem substitutions [13, 14] or insertions/deletions of one or more nucleotides.

Mutations affect the organismal fitness, and it is on these fitness differences that selection can act. Biologists offer no single definition of fitness [15] and theoreticians debate how it is best described [16]. We proceed with a simple definition, that absolute fitness describes the expected reproductive success of the organism. Next, and henceforth, we disregard the absolute definition and instead uses a relative fitness, ω, where fitness is normalized by the fitness of the fittest type, so $\omega \in [0, 1]$. Relative fitness is typically used since, in nearly all cases, only fitness differences matter to selection [15].

2.2 The fate of mutations is stochastic

Much insight can be derived from studying the deterministic differential equations that describes the growth behavior of competing species [17], but that approach cannot describe the important effects that stochasticity has in natural evolution. Instead, and of relevance for paper I-III, we will analyze the evolutionary dynamics in populations of constant size resulting from stochastic birth-death processes. Two birth-death processes are of special prominence: The Moran process and the Wright-Fisher process. In the following we simulate evolution using the Wright-Fisher process, which is as follows: Consider a population with constant size N. The population has at least two types, A and B, with fitnesses ω_A and ω_B, and initial frequencies $f_0(A)$ and $f_0(B)$. At each time step, all individuals reproduce with a probability proportional to their relative fitness, and thereafter, all individuals of the parent generation die. This birth-death behavior is similar of that of for instance annual plants.

To illustrate the behavior of the Wright-Fisher process, we start by considering the evolutionary dynamics when type A and B has the same fitness and no mutations occur. In figure 2.1 we follow 50 trajectories of $f(A)$ for various N and initial frequencies $f_0(A)$. We see that a special situation occurs whenever a type overtakes the population achieving so called fixation. At this point no further evolution happens until new mutations occur. As a result, without mutation, selection acts to purge populations of diversity. We can also observe that fluctuations are much more rapid when the population size is small, which result in both higher fixation probability and as shorter time to fixation. The random fluctuations (drift) in genetic variation also means that new exactly fitness neutral mutations can attain fixation.

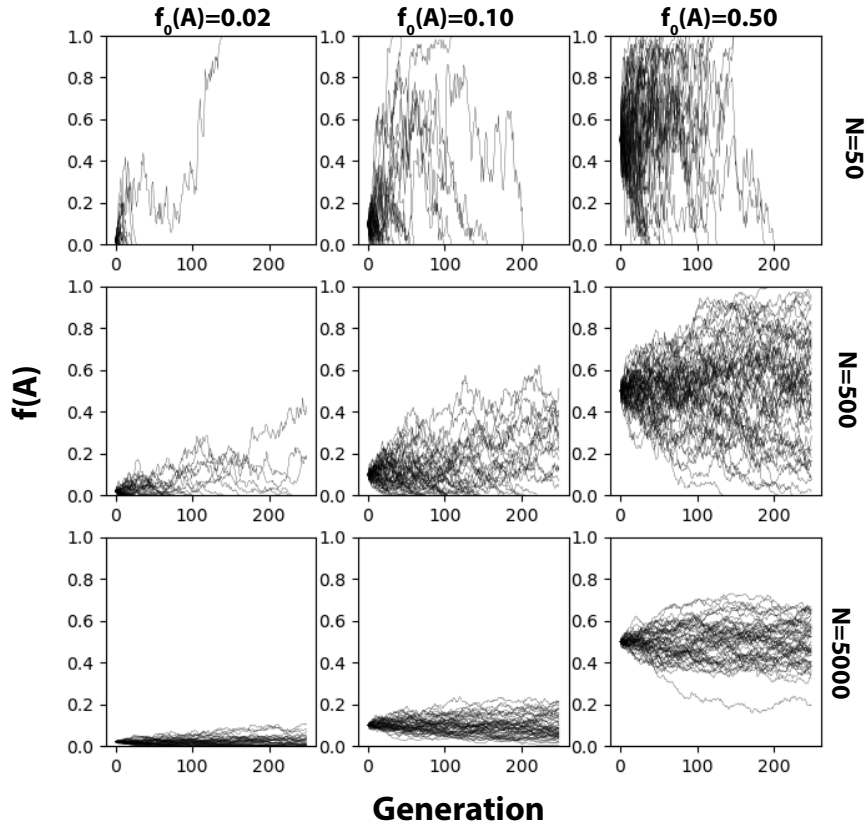


Figure 2.1: Stochastic simulations of constant size (N) populations evolving through 250 generations under the Wright-Fisher process. The initial population consists of $f_0(A) * N$ instances of type A and $(1 - f_0(A)) * N$ of type B, where f_0 is the initial frequency. Both types have the same fitness.

Next, we consider what happens when type A and B have different fitnesses. Of special interest is the situation where a novel mutant B arise in an otherwise homogeneous population of type A. It is useful to define the selection coefficient (eq. 2.1), s, which measures how much fitter B is compared to A:

$$s_{A \to B} = \frac{\omega_B}{\omega_A} - 1 \tag{2.1}$$

In figure 2.2, we follow possible trajectories of frequency $f(A)$ starting from one individual in populations of various sizes and with varying selection coefficients. We see that selection is not omnipotent — advantageous mutations can be lost due to genetic drift and slightly detrimental mutations can attain fixation.

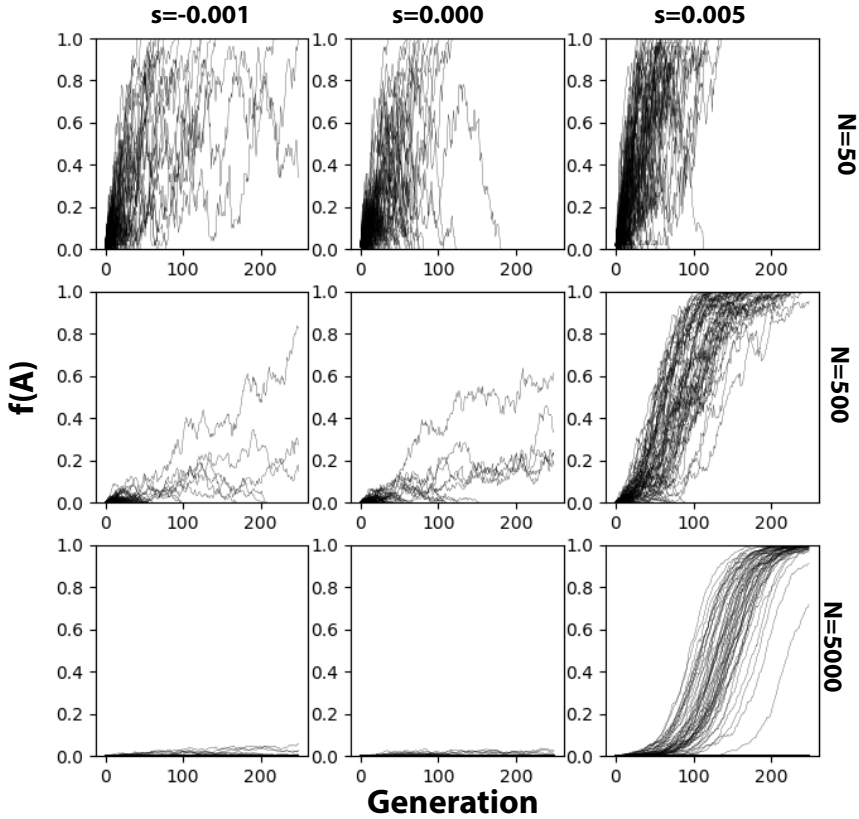


Figure 2.2: Stochastic simulations of constant size (N) populations evolving through 250 generations under the Wright-Fisher process with different selection coefficients (s). Every simulation start with one instance of the mutant (A) in a population of (N-1) individuals of type B

To understand the dynamics quantitatively, we simulated 50.000 independent trajectories and calculated the mean fixation probability for different population sizes and selection coefficients (Fig. 2.3). We see that a range of selection coefficients are nearly neutral (have fixation probabilities equal $1/N$) when $s \in [-1/2N, 1/2N]$. An exact analytical description of the expected fixation probability of new mutations under the Wright-Fisher model is not known. Motoo Kimura [18], derived a very good approximation (eq. 2.2), which is shown as lines in Fig. 2.3.

$$P_{A\to B}^{Kimura} = \frac{1-\exp(-2s)}{1-\exp(-2Ns)} \tag{2.2}$$

A slightly more accurate approximation was recently derived by Sella and Hirsh [19].

$$P_{A\to B}^{Sella\&Hirsh} = \frac{1-(\omega_A/\omega_B)^2}{1-(\omega_A/\omega_B)^{2N}} \tag{2.3}$$

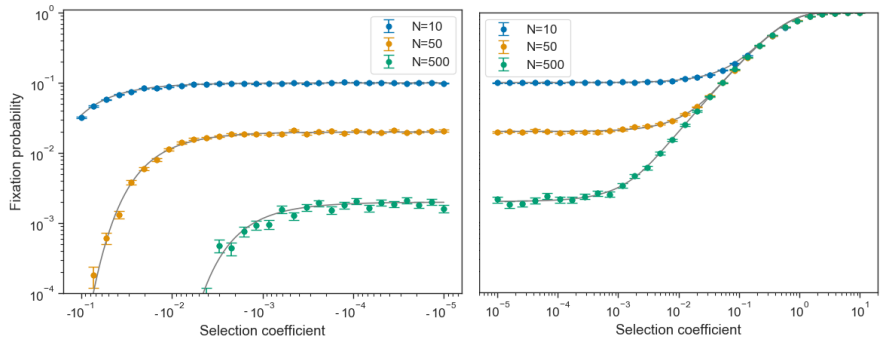


Figure 2.3: The relation between fixation probability and selection coefficient for population size 10, 50, and 500. Each dot represents the mean fixation probability across 20.000 Wright-Fisher simulations. The error bar represents the standard error of the mean. The grey lines shows Kimura's fixation probability equation.

The Moran process is identical to the Wright-Fisher birth-death process with the exception that it models overlapping generations, as only one individual is selected for reproduction and one for death at each time step. This result in genetic drift that is twice as strong as under the Wright-Fisher process, but qualitatively the dynamics are similar.

Throughout this work, we assume that mutation-fixation events occur according to the Wright-Fisher process in a constant environment as described by Kimura's equation. Although the Wright-Fisher process is robust to changes in the birth-death process, it does not capture all aspects of reality. In nature, population sizes and selection pressures fluctuate, resulting in off-equilibrium effects, which broadens the range of neutrality and modifies the fixation probability in specific ways depending on the exact fluctuation behavior [20]. By applying Kimura's fixation equation, we also assume that new mutations arise in otherwise monomorphic populations. This is only true under conditions where the mutation rate is low and selection is strong. It is known that heterogeneity requires special considerations [21, 22]. Likewise, special considerations must be taken when the population is subdivided [23]. The degree by which subdivision, heterogeneity, and fluctuations affect the mean behavior of evolution in general, is unclear.

2.3 The equilibrium distribution of states

For a given element, be it a particular gene, a single position in a protein, or a higher-level phenotype, one can describe it as situated in a finite state-space. For proteins, John Maynard Smith, pictured the vast sequence-space and hypothesized that all proteins were part of the same continuous network, each sequence connected to all other through mutational paths [24]. The transitions between states can be described by a Markov process with rate matrix, $\mathbf{Q}$:

$$\mathbf{Q} = \begin{pmatrix} -u_{AB}P_{AB} & u_{AB}P_{AB} \\ u_{BA}P_{BA} & -u_{BA}P_{BA} \end{pmatrix} \tag{2.4}$$

where u_{AB} represents the proposal rate of mutation between state A and B, while P_{AB} is the fixation probability of the mutation. Assuming that all states are connected (all states can mutate to all other states through some, possibly multi-state, path) and that detailed-balance is satisfied[2], the Markov process has a unique equilibrium distribution, which assuming that $s \ll 1$ in eq. 2.2 is given by

$$\pi_i = \frac{\theta_i \exp(2N\omega_i)}{\sum_j \theta_j \exp(2N\omega_j)} \tag{2.5}$$

where θ_i is the degeneracy of the state, N is the population size and ω is the relative fitness of the state [19, 26]. Sella and Hirsh noted that this is analogous to the distribution of states (conformation) in a thermodynamic ensemble, which is described by the Boltzmann distribution [19], but where physical ensembles depends on energy and temperature, evolutionary ensembles depends on fitness and drift.

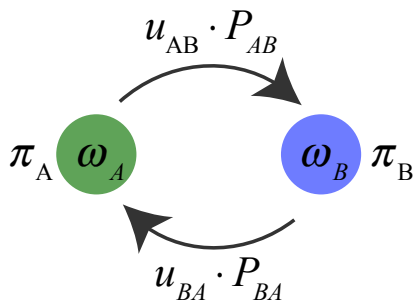


Figure 2.4: A simple Markov process where transitions happen between two states. Evolutionary transition rates are determined by the mutation proposal rate (u) and the fixation probability (P). The equilibrium frequency of each state is given by 2.5.

The flux between any two states is $\Phi_{AB} = \pi_A Q_{AB}$ and the exchangeability is $r_{AB} = Q_{AB}/\pi_B$. In paper II, we seek to reproduce the exchangeability and stationary frequencies between amino acids in evolution.

2.4 The rate of evolution

The rate of evolution is of special interest and again is only examined under equilibrium conditions. Also here Markov processes provide an apt mathematic framework

[2]The Moran birth-death process with mutation and selection satisfies the detailed balance condition $pi_A Q_A B = pi_B Q_B A$ for all A, B, while the Wright-Fisher process only satisfies it in certain limits [25].

to describe it. The rate of a Markov process is the sum of fluxes leaving all states. The flux out of one state is simply it equilibrium frequency multiplied by its rate to all other states, where rate in evolutionary dynamics is described as the product of mutation rate and the fixation rate.

$$\mu = \sum_i \sum_{j \neq i} \pi_i Q_{i \to j} \tag{2.6}$$

To get an intuitive understanding of the trade-off between selection pressure ($S = Ns$) and rate, we consider the simple situation illustration in Fig. 2.4. We set the mutation proposal probability equal in both directions $u = 1$ and determine the rate as given by eq. 2.6, 2.1, and 2.5. Figure 2.5 shows how the rate decreases as the selective pressure increases.

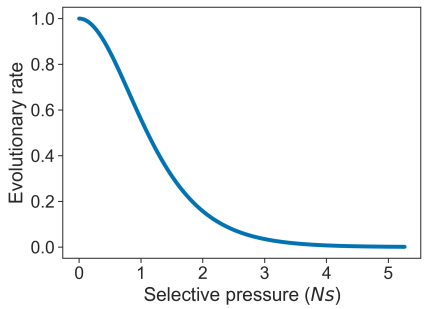


Figure 2.5: The rate of evolution at one site decreases with increasing selection pressure. The rate is calculated relative to the rate of neutral mutations by numerical evaluation of equation 2.6, 2.1, and 2.5 for the Markov process shown in fig. 2.4.

2.5 Inferring protein properties from sequence alignments

In the sections above, we have discussed amino acid frequencies, site-rates, and substitution rates between amino acid as if they were measurable quantities. They are not. No observable data exists, which directly describes them. Yet, as with the mass of an unobservable planet causing a slight wobble in the light from a distant star, these hidden variables can be inferred from other observations interpreted through models of the signal generating process. In this thesis work we have not attempted to improve on those statistical models but have used them as given. For this reason, we shall not detail the statistical and phylogenetic methods here, but refer to Felsenstein's book on phylogeny for that purpose [27]. Instead, we aim provide a sense for the high-level logic that goes into them.

Protein properties are often inferred in bioinformatics and phylogenetics using the Markov models described above. Assuming site-independency, a single site can be described by the rate matrix for site L, $\mathbf{Q}^L$ with the site-rate μ (eq. 2.6). At this,

already approximate level of description, the model contains 209 free parameters (190 exchangabilities and 19 stationary frequencies) per site in the modeled protein. Most sequence alignments do not contain sufficient amount of information to reasonably allow parameterization of that many parameters. Instead, further approximations are made. It is often assumed that a single substitution rate matrix ($\mathbf{Q}$) is relevant for all sites[3]. This matrix is typically scaled so that $\mu = 1$, which corresponds to that one time step yields 1 expected amino acid substitution per site.

A major variation in evolutionary behavior across sites, is the site-specific rate. Therefore, while maintaining an invariant relative substitution pattern, $\mathbf{Q}$ is typically scaled by a site-specific rate parameter μ^L, the likelihood of which is typically described by a gamma-distribution. Using site-specific rates, site-specific rate matrices are calculated as, $\mathbf{Q^L} = \mathbf{Q}\mu^L$. The probability matrix for substitution between any two amino acids over a given expected number of substitutions (thought of as proportional to time) is:

$$\mathbf{P} = \exp(\mathbf{Q}^L t) \tag{2.7}$$

the right term denotes the matrix exponential. For a given phylogenetic tree, T, describing the evolutionary history of the data (alignment of sequences), D, a set of site-specific rates μ, a substitution matrix, Q, a likelihood can then be calculated as:

$$\mathcal{L} = \prod_L \mathcal{L}(D|\mathbf{Q}, \mu^L, T) \tag{2.8}$$

where the product runs over sequence data for all sites, L, and $\mathcal{L}(D|\mathbf{Q}, \mu^L, T)$ is the likelihood of the data given the parameters. The unknown parameters, here the substitution matrix ($\mathbf{Q}$) site-rates (μ), and the tree (T), can then be inferred by finding the parameters, which maximize the likelihood of 2.8. More generally, the use of such maximum likelihood models is how protein properties are often inferred from sequence alignments, and methods are available to infer site-rates, substitution matrices, site-specific equation frequencies of amino acids, trees and more. Sensible Bayesian approaches, which takes prior probabilities of parameters into account has also been developed, and were used in the inference of rates in paper II [28].

We finally note that the sequence alignment is not observable, but should in principle be part of the parameter set that is being inferred [29]. Through this work, we have taken sequence alignments (typically generated with MAFFT [30]) as given point estimates.

[3]Given to the chemical and functional heterogeneity between sites in proteins, this assumption is clearly absurd. However, it works sufficiently well on average that it practice provides a good starting point for inferring phylogenies.

Chapter 3

The evolution of proteins

Life is the mode of action of proteins [1]

In this chapter, we focus specifically on the evolution of proteins. We discuss the many ways DNA mutations can change the sequence of a protein, distribution of fitness and energetic effects of mutations, and how the protein sequence relates to fitness and protein function.

3.1 Mutations can change proteins in various ways

Proteins are produced by translating RNA that has been transcribed from DNA. Each amino acid is encoded by 1 to 6 different codons (base triplet). Mutations in the DNA can change the protein sequence in several ways. Insertion and deletions (indels) of pieces of DNA that is not divisible by three result in frameshift mutations. They can change the entire protein sequence downstream of the mutation site and are typically detrimental to the function of the protein. Indel mutations, with a length divisible by three, change the length of the protein and might not be detrimental. Other mutations only change a single base or codon in the DNA. Due to the redundant nature of the genetic code, such mutations can be synonymous in that they encode the same amino acid. However, synonymous mutations are no guarantee that fitness is not affected, as some codons are translated more efficiently and/or accurately than others [32, 33]. Non-synonymous mutations result in substitutions to other amino acids (missense mutations) or might prematurely stop translation if resulting in a stop codon (nonsense mutation) or changes to the start codon. Finally, missense substitutions can also be introduced during protein synthesis and it has been estimated that around 18% of all average-length protein molecules contain at least one translational-missense error [34]. In this thesis work we have considered the fitness effects of missense mutations.

[1] Friedrich Engels according to William Henry Bragg [31]

3.2 What determines the fitness of a protein?

Each protein in an organism can be thought of as multiplicatively contributing to the organismal fitness. Returning to the relative fitness definition, each protein can, at most, contribute with $\omega_{max} = 1$ and at least with $\omega_{min} = 0$ if the protein is essential (non-essential proteins would have a $\omega_{min} > 0$). For brevity, we define the multiplicative fitness contribution of a protein to the organism, simply as *protein fitness*.

3.2.1 Fitness benefits

Protein fitness is the product of many biophysical properties. Ultimately, the benefit of a protein comes from its function, which is achieved by energetically favorable interactions. Through these interactions, proteins catalyze chemical reactions, bind ligands, transduce signals or perform other essential functions of life. With the possible exception of intrinsically disordered proteins, the folding of the amino acid chain is a prerequisite to correctly form a three-dimensional chemical environment compatible with function. It is therefore not surprising that many models of protein fitness have used the probability of the folded state as a proxy for protein fitness ([35]). It follows from statistical thermodynamics that a protein with folding free energy $\Delta G = G_{folded} - G_{unfolded}$ has a probability of being folded (P_{folded}) as shown in figure 3.1 and given by

$$P_{folded} = \omega_{folded} = \frac{1}{1 + \exp(\Delta G/RT)} \tag{3.1}$$

where R is the gas constant and T is the absolute temperature. For plots throughout this thesis work we used $RT = 0.6$ kcal/mol, corresponding to a temperature of $T = 300$K. In itself and for proteins that are active in their folded states, eq. 3.1 can be interpreted as multiplicatively contributing to protein fitness. The fraction folded fitness expression is also commonly used through the literature, in some cases in the form of a simple step-model (the threshold model) where $RT = \infty$ [35].

More realistically, it is not always sufficient for function that the protein is folded, as the thermal jitter of amino acids or loops results in many non-functional folded states. Given multiple competing states, the probability of one (active), is described by the Boltzmann distribution

$$P_{active} = \frac{\exp(E_{active}/RT)}{\sum_i \exp(E_i/RT)} \tag{3.2}$$

The "competition" between functional and alternative states has important consequences in natural proteins. For instance, it appears that evolution restricts the side-chain conformations of amino acids in binding-interfaces [36], possibly to avoid

non-specific interactions, but also to lower the entropy-cost of binding. The necessity to design against alternative states is perhaps most clearly demonstrated by the many cases of computational protein design, where reality found alternative ways to form interfaces [37], organize loops [38], organize catalytic site residues, than what was intended by the designer (for more examples see [39]).

3.2.2 Fitness costs

Drummond et al. demonstrated that another fitness function than ω_{folded} must be controlling the evolution of many, especially highly expressed proteins [40]. They saw empirically that the rate of evolution decreased with protein abundance (A) and hypothesized that the cytotoxicity burden of misfolded proteins imposed a significant selective pressure in evolution. Assuming the distribution of mutational effect on free energy is Gaussian ($\Delta\Delta G_{mean} = 1.0, \Delta\Delta G_{\sigma} = 1.7$kcal/mol) [41] and that 18% of all proteins carry a mistranslation error, it is clear that the majority of misfolded protein in the cell must be coming from translation errors almost independent of ΔG_{native} as shown in figure 3.2. Capturing these effects, they proposed the cytotoxicity fitness function:

$$\omega_{toxicity} = \exp\left(\frac{-cA}{1+\exp(-\Delta G/RT)}\right) \tag{3.3}$$

As shown in figure 3.1 and mathematically in supporting information of paper II, eq. 3.3 simply offsets ΔG in the fraction folded fitness function (eq. 3.1) by $\log(cA)$ when the protein is stable ($\Delta G/RT < -3$). They can therefore be used synonymously in many applications of evolutionary dynamics.

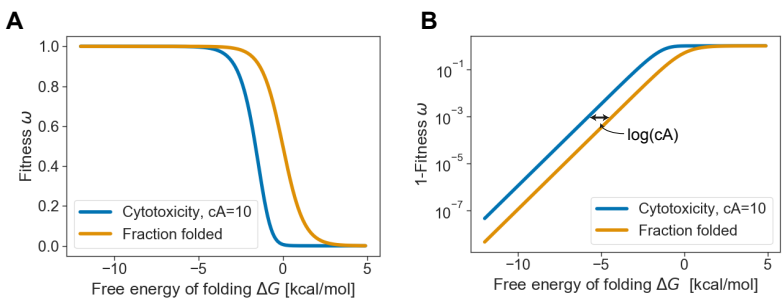


Figure 3.1: Two biophysical models of protein fitness. In one model, fitness is assumed proportional to the fraction of folded protein. In the other model, fitness is assumed to be a function of the amount (A), toxicity (c) and the fraction of unfolded protein. In the plot RT=1.6 and cA=10.

Serohijos et al. further analyzed the evolutionary dynamics under the cytotoxicity fitness function and found that the correlation between evolutionary rate and

protein abundance only holds when the mean effect of mutation becomes increasingly detrimental for more stable proteins [42]. Based on mutational effects recorded for real proteins in the ProTherm database [43], they showed that $\Delta\Delta G_{mean} = -0.13\,(\Delta G) + 0.23$ kcal/mol. This corresponds to a super-exponential growth of the sequence space with decreasing stability, which is also expected from informatics-based sequence studies [44].

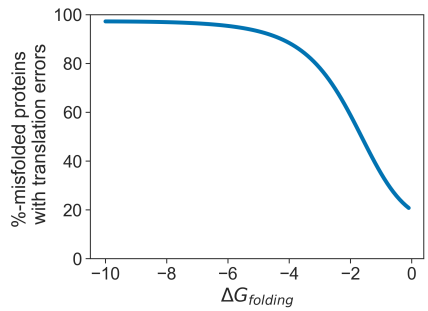


Figure 3.2: Most misfolded protein in cells is the result of translation error. The fraction of misfolded protein that results from translation error was calculated assuming that 18% (corresponding to an error rate of 10^{-4} and a protein with length 400) of all protein has one translation error and sampling mutational effects on δG from a Gaussian with parameters $\mu = 1.0, \sigma = 1.7 kcal/mol$ [41]. The x-axis shows the folding free energy of the unmutated native protein.

Besides misfolding cytotoxicity, proteins can also impose other fitness costs on organismal fitness. For instance, it appears that selection against misinteractions imposes a significant selective pressure and especially so in highly expressed proteins [45]. Levy et al. demonstrated that the surface of highly expressed proteins tends to be decorated with "non-sticky" amino acids [46]. In addition to changing their surface properties, proteins also use non-ideal secondary structures (irregular edge stands in β sheet) to avoid misinteractions (stand pairing) and aggregation with other proteins [47, 48]. Finally, the energetic cost involved in the synthesis of highly expressed proteins also imposes a fitness cost [49], possibly explaining the preferential use of ATP-cheap amino acids in highly expressed proteins [50].

3.2.3 Balancing benefits and cost leads to frustration

Although stability and misfolding-avoidance are aligned in terms of selective directionality (both work to increase protein stability), it is not always so. For instance, it appears often impossible to select an amino acid that is optimal for both stability and function. Shoichet et al. first hypothesized and demonstrated that significant improvements in stability could be gained by mutating catalytic or ligand binding residues in T4 lysozyme [51]. Similar effects have been determined in β-lactamase [52]. The function-stability trade-off is sufficiently strong that it allows prediction of functional sites from structure-based predictions of energetic frustration [53–56].

3.2.4 Fraction folded fitness functions results in marginal stability

Experimentally it is known that most proteins are marginally stable (the mean stability of protein in *E. coli* is -7.1 kcal/mol [57]). DePresto et al. took this to mean that the fitness of proteins does not follow the fraction folded fitness equation (eq. 3.1), but instead, that fitness had to decrease when protein stability increased beyond a certain point, arguing that a too stable protein would be harder to regulate (be protease-resistant) and have lower activity [58]. Goldstein et al. took a different approach in explaining the marginal stability of proteins. Using both lattice models [59, 60] and later more detailed models of protein energetics [61], they demonstrated that marginal stability arises directly from (i) a fraction folded fitness function, (ii) genetic drift, and (iii) the shape of the sequence space.

3.3 The effect of mutations

What happens when amino acid altering mutations arise in protein coding sequences? Here we examine the effect on protein fitness as well as on protein stability.

3.3.1 The distribution of fitness effects

Motoo Kimura made theoretical insights into the distribution of fitness effects. Seeing that the rate of evolution must depend on the rate of mutation per replication (u) the population size (N) and the rate of fixation (P_{fix}) the total rate of evolution (total number of mutations that fix) is $R = NuP_{fix}$. For neutral mutations where $s \in [-1/2N, 1/2N]$, we saw that the fixation probability, P_{fix}, was $1/N$ (eq. 2.2), and thus for those $R = u$. The fixation probability of detrimental mutations depend on the population size, but would not affect the observed rate of evolution, as they never fix (consider the sharp decrease in fixation probability in Fig. 2.2). The fixation probability of beneficial mutations also depend on the population size and would be predicted to affect the rate of evolution if significantly present in the distribution of fitness effects (DFE). However, if beneficial mutations are very rare compared to neutral mutation, and if the fraction of neutral mutations is constant, Kimura's theory predicts that the rate of evolution is constant. This is consistent with the observed constant rate of evolution and provided the theoretical explanation for the molecular clock [8], which lies at the foundation of molecular phylogenetics.

Experimental data are now available from several organisms and genes, which support Kimura's theoretical prediction of the DFE. Mutational studies of Hsp90 (a yeast chaperone) [62], of an RNA virus [63] (see Fig. 3.3), of β-lactamase in *E. coli* [64] all demonstrates that the vast majority of mutations are either neutral or detrimental.

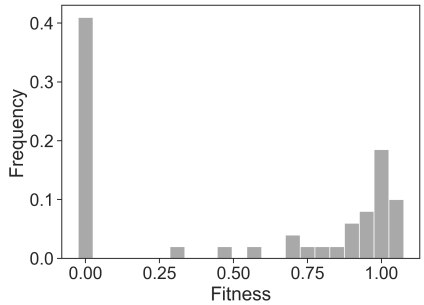


Figure 3.3: Distribution of fitness effect for random mutations in an RNA virus. A large proportion of mutational effects are detrimental (fitness=0) or neutral (fitness=1). Only, very few mutations are beneficial (fitness > 1) Data adapted from Sanjuan et al. [63].

3.3.2 The distribution of stability effects

Given that all positions in proteins are under selective pressure to contribute to the fold-stability, we next consider the distribution of energetic effects of mutations (DEE). Tokuriki et al. studied the DEE using researcher-recorded mutational effect of which thousands have are stored in the ProTherm database [65]. They found that the distribution did not appear to depend on the fold, but was significantly different for buried and exposed sites. The stability effects at surface exposed sites were generally lower than mutating a buried site. The distribution of energetic effects is shown in figure 3.4. Wylie and Shakhnovich showed that the DEE could explain the DFE in stochastic simulations of populations that allowed for genetic drift and polyclonality (high mutational load) [41].

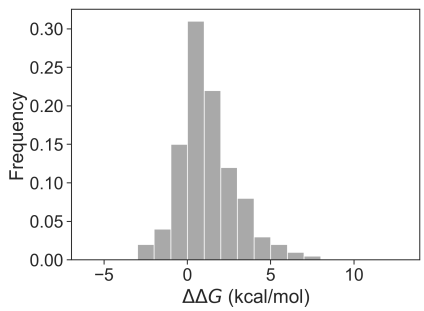


Figure 3.4: Distribution of energetic effect of amino acid substitutions in proteins. Data adapted from ProTherm [43] as summarized by Tokuriki et al. [65].

3.4 Energetic and fitness epistasis

In real proteins, amino acids interact. That is, the energy contribution of a site to the total energy of the protein, depends on the exact chemical environment that the site sits in. The non-additive interaction between different mutations is called epistasis. The effect is sufficiently strong that covariation between amino acids is observed in sequence alignments of homologous proteins and allows for the prediction of neighboring positions in protein structure (see for example [66, 67]). Curiously, these methods work without considering the phylogenetic relationship between sequences and does not attempt to model evolution, which could in principle lead to spurious covariation signals [68]. Instead, they identify coupled positions by finding the maximum likelihood of energetic parameters in a global statistical model that describes the sequence alignment. Specifically, the energy of a sequence is determined as

$$E_{a_1,a_2,\ldots a_n} = \sum_{i<j} \mathbf{w}_{ij}(a_i, a_j) + \sum_i \mathbf{v}_i(a_i) \tag{3.4}$$

where $\mathbf{w}_{ij}$ is an $20x20$ matrix representing the energetic couplings between site i and j, and $\mathbf{v}_i$ is a 20 long vector with non-coupled energies for amino acids at site i. The energy of the sequence is then converted to a probability through the Boltzmann distribution equation $P = \exp(E)/Z$, where Z is likelihood normalized over all sequences. The log-probability of the alignment, which is what is being maximized, is computed as the sum over all sequence log-likelihoods.

Fitness epistasis might also arise without direct interaction between residues, as can be seen in figure 3.5 by combining eq. 3.1 and 2.1 and assuming that $\Delta G_i \ll 0$ and $\Delta G_j \ll 0$.

$$s = \exp \Delta G/RT(1 - \exp(\Delta\Delta G/RT)) \tag{3.5}$$

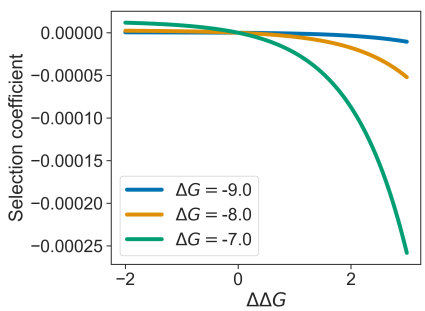


Figure 3.5: The selection coefficient of one mutation resulting in $\Delta\Delta G$ depends on the energy of sequence ΔG it occurs in.

In effect, all positions interact with all other positions by being coupled through the total stability of the protein. However, as ΔG fluctuates much more rapidly than single positions mutates, this effect is unlikely to result in observable covariation between amino acids in homologous proteins. Instead, the coevolution between positions in protein originate from positions that are coupled through reciprocal sign epistasis [69] illustrated in Fig. 3.6.

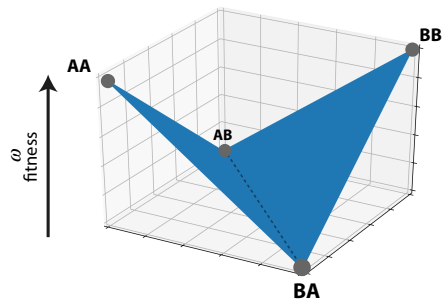


Figure 3.6: The fitness-sequence space of a binary two-site system exhibiting reciprocal sign-epistasis. The effect of combined mutations has the opposite effect of what would be expected from additivity of the individual mutations.

Talavera et al. [68] questioned whether the observed covariation could be the result of fitness epistasis between sites. Under the fitness-sequence landscape illustrated in Fig. 3.6 and the theory described in chapter 2 they found that observable epistasis would only arise under very high selective pressure, and that, as we also saw in Fig. 2.5, this would result in extremely slow evolutionary rates. They called the discrepancy between observed covariation and expected covariation, "the coevolution paradox". In paper II we address this paradox.

From a site-centric point of view, epistasis results in fluctuations of equilibrium frequencies of amino acids. Pollock and Goldstein [70] showed, using the evolutionary dynamics framework of chapter 2, that fluctuations follows certain patterns. Whenever mutations do not quickly revert [71], the environment becomes increasingly more compatible with the mutant identity – resulting in a type of entrenchment, where reversion becomes increasingly more unlikely. Real proteins also display entrenchment [72, 73].

3.5 Determinants of evolutionary rates in proteins

So far, we have considered general properties of protein evolution. We saw that fraction-folded fitness functions explained both the marginal stability of proteins and the abundance-rate correlation. Further, we saw that the distribution of energetic effects applied using a fraction folded fitness function was consistent with the distribution of fitness effects. In sum, we explained some of the key properties of proteins.

To get an even more detailed understanding of the evolutionary forces that act on proteins, we now turn to the determinants of the evolutionary rate at specific sites in proteins, as well as the substitution rate between different amino acids.

In section 2.4 we saw that the rate of evolution decreases as fitness differences between types becomes increasingly more pronounced. As such, the evolutionary rates provides a site-specific, although non-linear, glimpse into the strength of evolutionary pressure. What are the known determinants of site-sites (and thus evolutionary pressure) within proteins?

3.5.1 Structural determinants of rate variability

Starting from the distribution of energetic effects discussed in section 3.3.2 where Tawfik et al. found that core sites were less tolerable to mutations than surface sites, we can make the general prediction that surface sites evolve faster than core sites. Such a pattern was already noted by Perutz et al. in 1965 [74], and confirmed more generally in one of the early studies of structural rate determinants where Goldman et al. found that surface sites appears to evolve approximately twice as fast as buried sites [75]. Later, Worth et al. reviewed structural rate determinants, but on an atomically-resolved scale. While they confirmed the importance of solvent accessibility, they also noted that buried hydrogen bonds that stabilize the protein architecture, for instance by capping the α-helices or rigidifying loops, resulted in significant evolutionary constraints [76].

Several phenomenological descriptors have been developed to predict the evolutionary rate given a protein structure. The phenomenological models typically neglects the atomic-details like those identified by Worth et al. and instead starts from coarse structure descriptors. In line with the finding by Goldman et al., the relative surface exposure (RSA) was applied to investigate the dependency across thousands of yeast proteins [77] and an almost linear (although noisy) correlation was found. The weighted contact number (WCN), another phenomenological descriptor of evolutionary rates, has superseded the predictive performance of RSA (for an excellent review on the many papers using WCN, see [78]). The WCN is determined as the square distance weighted number of residues of all other residues in the protein. In effect, WCN emphasizes the local environment like RSA, but in effect it also accounts the overall shape of the protein. It appears likely that the additional performance of the WCN is just a consequence of the fact that many functional sites (with very high selective constraints and thus low rates) are situated in topologically concave regions of the protein structure.

Models that predict site-rates from the first principles of evolutionary dynamics theory as described in chapter 2 and biophysical models of protein stability have also been developed. Jiang et al. generated evolutionary trajectories for 38 natural proteins and compared the resulting site-variation to empirical site-variation [79]. The site-variation predicted by the model was poorly correlated ($r^2 = 0.16$) to the empirical

variation. An alternative approach was taken by Echave et al. [80] who combined the Markov chain formalism illustrated in figure 2.4 with a mean-field approximation [81] to generalize the stability effects of mutations conditioned on a single structure for site-rate prediction. Although incapable of generating evolutionary trajectories, this approach yielded site-rate correlations of $r^2 = 0.30$ on par with those calculated using structural predictors such as WCN .

In sum, both phenomenological and biophysical evolutionary dynamics-based models predict site-rates rather poorly. Why are first principle models not out-performing the coarse phenomenological models? One possibility is that the biophysical models of protein energetics are inaccurate, which indeed they are (see chapter 4), but as we show in paper III, this does not appear to be the reason. Another possibility is that rates, due to epistasis, fluctuate significantly over evolutionary time-scales, as paper II indicates, and that taking a single structure as representative for all sequence environments, is not a good approximation. We intend to study this in the future. A third option, is that non-structural (non-stability) selection pressures significantly affect rates. We believe this is less likely as functional sites only make up around 10% of all residues [82]

3.5.2 Functional determinants of site-rate variability

Functional sites, such as active site residues in enzymes and binding residues in ligand or protein-protein interactions, are under strong selective pressure, and therefore have severely depressed evolutionary rates [83]. The magnitude of the rate depression depends on the type of function — allosteric sites and non-obligate protein-protein interactions have for instance higher rates than catalytic residue, but see the review of Echave et al. (2016) for a more completely review on the effect of function-type on rate depression [78].

The prediction of functional sites from protein structures, is an important problem in biochemistry, as researchers often seek to understand the mechanism of novel proteins. Although the rate-depression resulting from functional selection constraints, is sufficiently high and consistent to enable prediction of functional sites from rates alone [84, 85], it is not a perfect measure, as stability constraints also depresses site-rates. In paper III, we develop a method to predict functional sites from rates, while taking the expect rate-depression from stability into account.

3.5.3 Determinants of the substitution rate between amino acids

Averaging across sites, another measure that informs on general evolutionary pressures becomes apparent: The rate by which different amino acids substitution with one another. This substitution behavior is cornerstone in bioinformatic and phylogeny and has been phenomenologically described with two different approaches. Probability matrices such as the BLOSUM [86] have been highly useful for sequence alignments

and homology search, but are not based on an underlying model of evolution [87]. In contrast, models that explicitly describe the process of evolution have been useful for comparative sequence analysis. These models describe the substitution pattern as an aggregated Markov process where sites evolve independently from each other and with the same substitution rates that depends on the identity of the current amino acid (WAG matrix [88]) and possibly with a site-specific rate parameter (LG matrix [89]). To understand the origin of the variation in substitution rates, several studies have established statistical correlations between biophysical amino acid descriptors (such as secondary structure propensity, charge, codon count ect.), and the empirical rate variation [90–92]. It has also been demonstrated, that significantly better description of the evolutionary process can be achieved if multiple secondary structure-, or site-rate-dependent matrices are used [93, 94]. Both approaches have yielded an understanding of the factors important, but it remains an outstanding question which underlying fitness pressures causes these substitution patterns. We investigate this is in paper III.

3.6 The evolution of protein function and antibodies

It is curious to note that just a small subset of protein folds are so innovative that they can encompass the majority of all enzyme activities. Similarly, that our immune system is capable of recognizing almost any foreign antigen, is thanks to the innovability of just a single fold – the immunoglobulin fold (the subject of paper V to VIII). Dellus-Gur et al. hypothesized that the key to innovability is modularity (in their words, "polarity"): That the functional site is composed of flexible loops attached to independent stable scaffolding elements [95]. Nowhere is modularity more evident than in the immune molecules, here illustrated by antibodies (the design and prediction of which is pursued in paper V-VIII), as they evolve under constant evolutionary pressure to be universally modular.

The adaptive immune system of vertebrates, which arose approximate 500 million years ago [96], is one of the most astounding molecular marvels of evolution. From five, almost universally modular parts, each encoded by less than 100 genes, the immune system is capable of generating almost 10^{10} different molecules [97]. Each part (a gene segment) encodes a specific structural element in the immunoglobulin protein as illustrated in figure 3.7. Function (antigen-binding) typically derives from a limited portion of the antibody structure called the complementarity determining region (CDR), which spans two interacting immunoglobulin domains (the light and the heavy chain) the relative relative orientation of which can vary based on the residues in the domain interface. Each immunoglobulin domain presents three CDR loops. The first two loops are encoded by a single gene segment, while the last loop derives from an error-prone recombination event between either two or three gene segments. Despite their sequence diversity, the backbone conformation of most CDRs, except

heavy chain CDR 3, falls into a handful of distinct canonical conformations. For instance, light chain CDR 1 adopts only approximately five conformation across all 1,000 solved antibody structures [98]. The recombination is not done with a specific target "in mind" but instead, binding antibodies are selected from a repertoire of germline antibodies [99]. James et al. [100] found that some germline antibodies display considerable conformational diversity, allowing for initial multi-specificity. That conformational dynamism enables multi-specificity is an important principle in the evolution of function and has also been demonstrated for enzymes [101], where a duplication event can free the redundant copy from selection on the native function and allow further optimization for new functions.

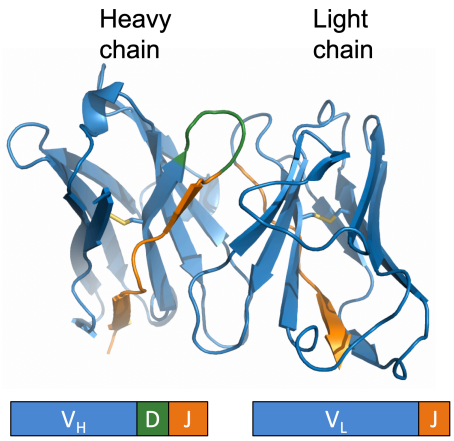


Figure 3.7: The structurally modular antibody. The variable domain of antibodies is created by genetic recombination of five different gene segments. The light chain is made through the recombination of V and J gene segments, while the heavy chain is created by recombining V, D and J gene segments. Constant domains are not shown. This figure, originally created by the present author for a presentation at the Weizmann Institute of Science, was reproduced previously in the thesis work of Lapidoth [102].

Chapter 4

Prediction and design of protein structures

What I cannot create, I do not understand [1]

Protein design offers immense opportunities for medicine and the chemical industry, and also provides the ultimate test of our understanding of molecular driving forces and the principles that shape natural protein diversity. Protein structure prediction serves as an essential tool in protein design and is also an invaluable tool in biochemistry. In this chapter, we discuss the mapping between protein structure and energy, how such energy functions are derived, their shortcomings, and how the dynamics of natural evolution impacts the work of protein design and structure prediction.

4.1 Macromolecular energy functions

An energy function is a mathematical function that, for a given set of 3D coordinates of a molecule, computes its corresponding energy. Throughout this thesis work, we have used the energy function of the macromolecular modeling suite Rosetta [103]. The energy function is necessarily approximate, as the underlying physics is quantum mechanical and currently intractable to compute for design applications. Instead, Rosetta computes the energy as a linear combination of terms, that together approximates the energy of the structure in question,

$$E_{total} = \sum_i w_i E_i(\Theta, aa) \tag{4.1}$$

[1] Richard Feynman

where w is the weight of each energy term, Θ the coordinates of the structure, and *aa* the sequence. The energy function is optimized against various structure prediction and sequence design tasks, which we discuss in the following.

4.1.1 Parameterizing the energy function for structure prediction

Part of the energy function is parameterized following Anfinsen's thermodynamic principle of protein folding [104]: The native conformation of the protein attains the conformation where the Gibbs free energy is the lowest[2]. That proteins fold this way, allows prediction of protein structures knowing only their sequence. The principle also provides a target function for energy function parameterization: The weights of the energy terms and parameterization of atoms and amino acids, should be such that the energy function, when combined with sampling of structural degrees of freedom, reproduce the atomic-details of known protein structures (side-chain packing, hydrogen bonds, van der Waals interactions, and more) and identifies the correct structure from a set of decoys [106]. Sippl took the thermodynamic principle one step further and posited that not only would the conformational ensemble of the protein itself be Boltzmann distributed, but so would the distribution of specific sub-geometries observed in protein structures [107]. This was later justified from a Bayesian point of view by Simons et al. [108] and formalized by Hamelryck et al. [109]. Today, many of the terms in Rosetta are individually parameterized from distributions of geometries in protein structures. For instance, the backbone potential in Rosetta derives from amino acid dependent probability distributions of backbone angles, $P(\phi, \psi|aa)$, based on approximately half a million high-resolution protein structures, which are then converted to an energy potential via the inverted Boltzmann relation, $E = -\ln(P)$. Other interactions types such as Van der Waals have a more physics-based origin, applying a Lennard-Jones 6-12 potential with further approximations for computational efficiency [103].

4.1.2 Parameterizing the energy function for protein design

Another part of the energy function is parameterized based on assumptions about the evolutionary process and is necessary for protein design. The necessity arises since there, contrary to structure prediction, is no guarantee that the target structure will fold to any specific conformation. Instead, it becomes necessary to engineer an energy gap such that the target structure is energetically well-separated from all alternative conformations, and thus highly populated according to the Boltzmann distribution.

[2]We note that this is not an obvious consequence of physics or for that sake evolution. Proteins could also have been under kinetic control, such that native states are not in the global energetic minimum. Interestingly, however, Govinarajan and Goldstein demonstrated with simple lattice models that even if protein folding is under kinetic control, the native state will most often also be the global minimum [105]

Sequence design strategies to avoid alternative states are called "negative design" and spans both heuristic and explicit strategies [39]. The simplest negative design strategy, built directly into the energy function of Rosetta, is to approximate the energy of the unfolded state as a sum of amino acid-specific reference energies over all positions in the sequence. If accurate, this ensures the encoding of an energy gap between the target and the unfolded states. To ensure an energy gap to alternative folded conformations other strategies must be employed (see next section). To parameterize the reference energies of the energy function, it is also assumed that the amino acid frequencies at positions in protein alignments are distributed according to the Boltzmann distribution [110]. This assumption traces back to work by Steipe et al. [111] who, inspired by Sippl [107], applied the Boltzmann assumption to approximate the stability effects of mutations in antibodies. However, as should be clear from chapter 2 and 3 or simply by comparison of the Boltzmann distribution (eq. 3.2) and the distribution of genotypes given fitnesses (eq. 2.5), the assumption is not strictly correct. The consequence of this and a possible route for improvement is explored in Paper I.

4.2 Designing energy gaps with imperfect energy functions

In practice and in theory, particular design and structure prediction problems are particularly hard. One can think of the difficultly as being dependent on how hard the energy gap is to design or realize. Problems with small energy gaps achieved through interactions that are hard to measure accurately, are the hardest. Somewhat counterintuitively, this makes the design (prediction) of globular protein structures with large hydrophobic cores and large energy gaps to alternative conformations an "easy" problem, while the design (prediction) of loop conformations or the organization of active site residues is far more difficult to get right [39]. In the following we review strategies used to overcome difficult-to-design energy gaps.

4.2.1 Idealized protein design makes alternative states unlikely

Historically, protein design has been undertaken starting from the protein folds of nature. As we saw in section 3.2.4, there are no selection pressures that will optimize thermostability of proteins beyond what is necessary to achieve function and minimize cytotoxicity, and consequently most protein are marginally stable or "non-idealized". This property is also reflected in their structures, which often have holes or other structural defects, which encourage repacking into alternative conformations. One recent approach has been to disregard the nature-given scaffolds and to design idealized protein structures. This has resulted in designed proteins with stabilities far exceeding that of any natural protein, remaining folded at 95°C in 7 M guanidine hydrochloride (a denaturant) [112]. More recently, this approach enabled the generation of barrel-shape

topologies that precisely fit the geometries of target small-molecules [113, 114]. However, for understanding the properties of natural protein structures or reengineering them for new purposes, protein ideality cannot be the way forward.

4.2.2 Symmetrical design multiplies the energy landscape and leads to higher separation of states

Another approach, which has been immensely successful, has been to design symmetrical protein structures, such as icosahedral capsids. The multivalency of symmetrical protein interactions effectively results in a multiplication of the energy landscape, resulting in a far better separation of states. In paper VII, we review the biophysical principles of protein self-assembly in detail [115]. To illustrate, we refer to figure 4.1, which shows the energy landscape of association of subunits of the icosahedral capsid, Lumazine synthase, as mapped out by a global docking simulation with Rosetta. Figure 4.1A shows that the dimeric and trimeric interface are not well-separated from alternative orientations when forming independently. However, when forming across the interface of pentameric units, the energy gap is significant.

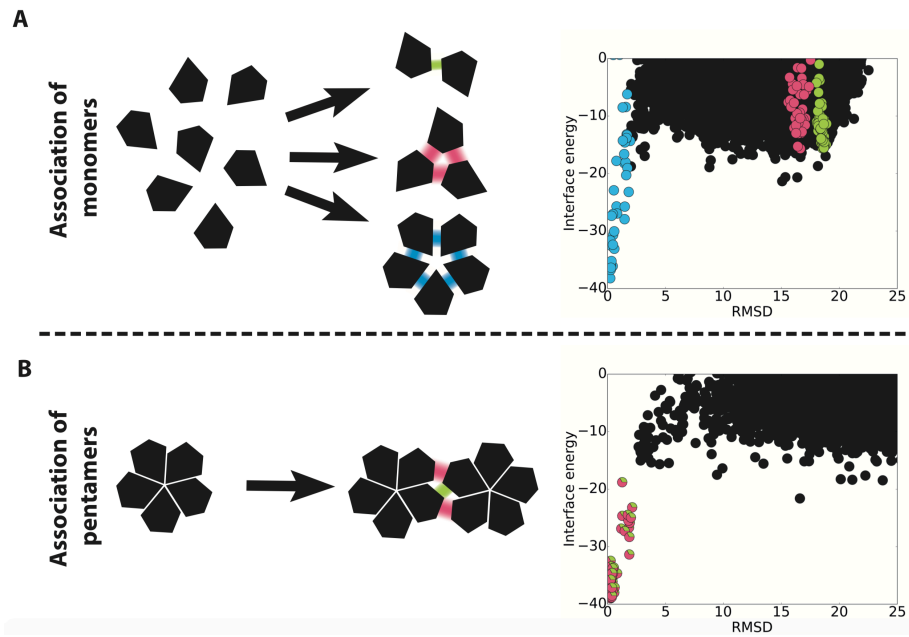


Figure 4.1: Energy landscape of Lumazine synthase, a protein that form capsids with icosahedral symmetry. Global docking simulations using Rosetta was done of either (A) monomeric subunits or (B) correctly formed pentameric subunits. Root-mean-square deviations (RMSD) to native dimers formed across the pentameric interface were calculated for all generated docking conformations. Conformations within 2 Å of the native pentamer, trimer, and dimer orientation are colored in blue, red, and green respectively. Correctly formed pentameric dimers (<2 Å) are colored in split red/green. The energy landscape was only funneled for the tri- and dimeric interface when formed simultaneously across pentameric subunits.

4.2.3 Explicit consideration of alternative states

Enumerating all alternative states is intractable except for the simplest of systems, as the conformational space increases exponentially with the number of degrees of freedom (rotatable bounds). However for simple systems such as helical bundles some progress has been made [116]. In situations, where the alternative states can be isolated to different (enumerable) alternative interactions, explicit negative design has been used to design orthologous pairs of both de novo and natural proteins [117, 118].

4.2.4 Heuristic design against alternative states

Implicit sequence design against alternative states can be conducted in multiple ways. For protein-protein interface design, one approach has been to ensure that the interacting residues would occupy high-probability constellations, reducing the conformational plasticity of the interacting surface patch [119, 120]. More generally, for design of surface and core positions, with no particular function other than ensuring structural stability, another approach is to limit the amino acid choices to what is generally observed in natural proteins for positions of comparable surface exposure and secondary structure (see for instance [121]). In paper IV, VI and VIII, we took a more quantitative approach and, following previous work [111, 122], we constructed alignments of sequences that encoded a similar structure as the target for sequence design (Fig. 4.2), and used log-odds-ratios of the observed amino acid frequencies compared to background frequencies to bias the energy function and to entirely restrict amino acids available for design to those observed. In paper I, we question whether the log-odds-ratios describe the underlying connection between observed frequencies and folding energetics. Nonetheless, this approach has now been successful also beyond the conditionally modular antibody scaffold where it has been applied to improve thermostability [123], design TIM-barrels [102, 124], and protein-protein dimers [118].

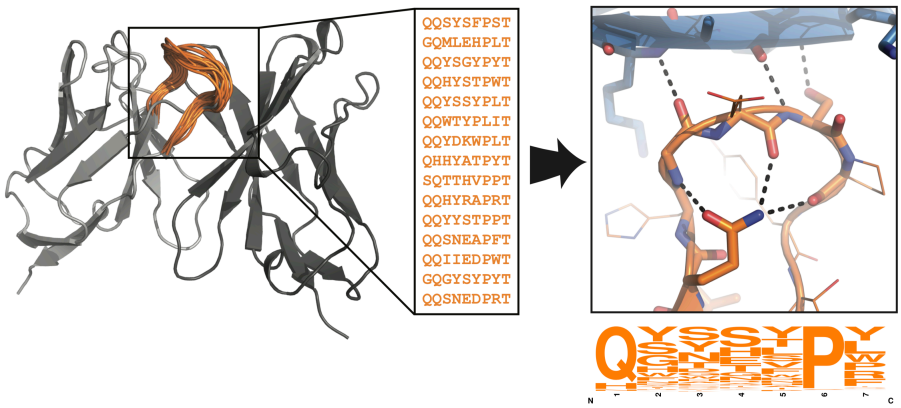


Figure 4.2: Evolution-based heuristic negative design approach for designing an antibody loop. Since more than a 1000 antibody structures have been deposited in the protein databank, and since most antibody loops have distinct canonical conformations, the allowable sequence diversity at each position in an antibody loop can be gauged accurately. This in turn can be converted into a position specific bias function, which steers computational design away from sequences with many unwanted alternative states (not shown). In this example, the naive energy function has a tendency to insert a hydrophobic amino acid instead of the highly geometry constraining glutamine at the first position in the loop.

4.3 Guiding sampling of degree of freedom in structure and sequence space

4.3.1 Vast but discrete sequence-structure spaces

Both in structure prediction and protein design, one must sample a space of conformations and/or sequences to find low energy solutions. The larger the sequence-structure space is, the more difficult it becomes to find the best solution and the larger the risk becomes that the energy function will make mistakes that will be detrimental to protein stability or obscure the energy gap, which should be observed for confident structure prediction. In section 4.2.4, we discussed ways to limit the sequence space based on observed sequences. Similar approaches have been taken in de novo structure prediction, where short sequence-similar structure-fragments with length 3 or 9 residues from known proteins have been used to limit and guide conformational sampling [108]. Today, for de novo structure prediction, the search space is typically even further restricted based on residue-residue contact constraints derived from statistical models over sequence alignments, as discussed in section 3.4. At the residue level the search space can also be restricted and discretized. Due to the backbone dependent intra-residue energetics, large barriers exists between different amino acid side-chain conformations [125]. When discretized to high-probability peaks in the corresponding likelihood functions, inferred from observed structures, these conformations are called rotamers [126].

For antibodies in particular, the guiding of conformational search can be extremely specific to the immunoglobulin fold because of the vast number of solved

antibody structures. In paper V, we disregard the 3/9-fragment assembly approach and instead use fragments corresponding to the splice points of gene segment recombination (see section 3.7). Additionally, the large number of known structures also allows the sampling of rigid-body orientation between the light and heavy chain from a discretized library. It is worth noting that the approach is dissimilar to traditional homology modeling, as sequence identity is not used to guide the selection of structure fragments.

4.3.2 Search methods

Given a set of possible geometries (backbones conformations, rotamers, or rigid-body orientations between subunits) or amino acid identities, the space must be searched with the goal of finding the minimum energy sequence, structure or both. The discrete nature of the search space lends itself well to a Monte Carlo simulated annealing (MCSA) optimization approach both for sequence design [127] and structure prediction [108]. In this approach, changes are accepted with $\min(1, P(x_{i+1})/P(x_i))$ where P is the Boltzmann factor $\exp(E(\Phi, aa)/T)$. This approach reproduces the Boltzmann distribution for any given temperature of the system [128].

The structure space is however continuous in reality, so often, between the discrete moves with fragments, sequence or rigid-body orientations, continuous exploration of the structure space is pursued using gradient-based minimization methods. At this point the protein structure can be modelled at different levels of granularity. Freezing bond length and minimizing the protein only in torsion space is highly efficient as it significantly reduces dimensionality [129], however the speed comes at a cost accuracy, which might obscure the energy gap in native proteins [130].

Chapter 5

Summary of thesis work

In this chapter, we summarize the findings of the papers.

5.1 Paper I

In paper I, we sought to predict stability effects using only the evolutionary sequence diversity of protein alignments as a source of information. The ability to predict the effect mutations have on protein thermostability is important in protein engineering as it correlates with longer shelf-life, higher operating temperatures (and thus higher activity flux in enzymes), longer in vivo half-life, and provides a more robust starting point for in vitro evolution or engineering new functions into a protein scaffold. Prediction of stability effects is also important medically, as it can be used to predict disease variants.

Methods that relate protein stability to observed amino acid frequencies have a long tradition in protein engineering, and most methods build on work done by Steipe et al. in 1994 who, with no basis in evolutionary dynamics theory, assumed that the relationship between protein stability and amino acid frequencies would be the Boltzmann distribution. This assumption is today embedded in a popular engineering method called consensus design, in computational protein design methods (including paper VI and VIII), and is applied when optimizing the energy functions. In paper I, we took a different approach and tried to connect amino acid frequencies to protein stability using the first-principles theory of evolutionary dynamics and explicit assumptions about the connection between protein stability and fitness. Several important conclusions were made.

First, we showed that the Boltzmann distribution assumption can be understood within the theory of evolutionary dynamics. In fact, it gives rise to the same relationship between stability and frequency (albeit with a scale factor) when protein fitness is directly proportional to protein stability. However, it is highly implausible that this would be the case. Instead, reviewing the literature, it appears far more likely that

the fitness-function over protein stability is sigmoidal, corresponding to fitness being proportional to the fraction of folded protein (and anti-correlated with the amount of unfolded protein).

Second, using evolutionary dynamics simulations based on the distribution of mutational effects in real proteins and the fraction-folded fitness function, we showed that the Boltzmann assumption still provided a good approximation and a perfect ranking of the energetics *within* sites. We were further able to rationalize its reasonable performance analytically. However, when comparing stability effects *between* sites, the Boltzmann assumption does not guarantee correct ranking between amino acid choices, as the evolutionary dynamical partition function is expected to vary.

Third, we provided an approximation of the evolutionary dynamical partition function, which tentatively, appears to improve detection of stabilizing mutations within proteins. However, more work needs to be done to establish statistical significance.

5.2 Paper II

The evolutionary behavior of proteins is of key interest to several fields including bioinformatics, phylogenetics and biochemistry. In this manuscript, we develop a novel first-principle biophysical model of protein evolution that describes the fitness of proteins according to their predicted thermodynamic stability. We apply the model to answer three unanswered questions: (i) Which fitness pressures control the observed patterns of amino acid substitutions and (ii) the site-specific rate of substitution? (iii) Can a thermodynamic-stability null-model of evolutionary rates identify sites under functional fitness constraints?

Today, bioinformatics and phylogenetics rely on empirical descriptions of the substitution behavior as summarized in matrices such as BLOSUM or LG, which describe evolution of all positions in proteins identically. A considerable push has been made to develop a small set of empirical matrices that take the local environment of each site (secondary structure for instance) into account. Ultimately, such matrices should depend on the exact chemical environment and evolutionary dynamics. Our model does exactly that and further allow us, for the first time, to investigate how much of the mean substitution behavior that can be explained by thermodynamics alone (at least 65%).

At a site-specific level, we investigated how thermodynamic fitness constraints affect the site-specific rate of evolution in proteins, which is critically important in phylogenetic inference. We find that our model is far better than current structure-based methods in predicting conserved sites. This property allows us to test a central unanswered hypothesis recently proposed by Echave, Spielman and Wilke [**?**], namely, that discrepancy between empirical rates and rates predicted with a thermodynamic null-model of evolution could be predictive of functional sites. We validate

this hypothesis, showing that the metric can predict up to twice as many functional sites as empirical rates alone. Structure-based prediction of functional sites is an important tool in biochemistry as illustrated by the 1200 citations accumulated by the ConSurf server, which only applies empirical rates for prediction of functional sites.

Our findings could be of interest for scientists in evolutionary biology, protein science and bioinformatics. We provide a model that gives insight into the fitness pressures that control the evolution of proteins, show how protein energetics influences natural sequence diversity and provide a first-principle approach to predict functional sites in proteins.

5.3 Paper III

In silico simulations of sequence evolution on protein structure has provided several important insights concerning properties of real protein evolution. So far, most in silico simulators of protein evolution have been based on lattice models of proteins or simple contact-based potentials that do not model the atomic-details of interactions. It is possible that all-atom models of protein evolution could reveal new properties or help gauge the magnitude of various effects. In this manuscript we establish one such model.

Following previous work, we model protein fitness as a sigmoidal function of the free energy of folding

$$\omega = \frac{1}{1 + \exp(\Delta G/RT + O)} \tag{5.1}$$

where O offsets the half-maximal fitness point of the function. If the fraction of folded protein is the only determinant of fitness $O = 0$. However, other selection pressure pressures, including the toxicity of misfolded proteins, result in $O > 0$, and thus require more stability for equal fitness. Additionally, the strength of mutational drift, determined by the population size, has the same effect, as it offsets the distribution of sampled protein stabilities by a factor proportional to $-\log(N)$ [41, 61]. Here we investigate how the offset of sigmoidal fitness functions influence the behavior of protein evolution.

We began by implementing an evolutionary dynamics framework within the macro molecular modeling suite Rosetta. Mutations were proposed at the nucleotide level, structure and energetic effects were evaluated using the established structural modeling tools in Rosetta, fitness was calculated with a sigmoidal function, and the fate of mutations determined according to Kimura's fixation probability. Using this framework, we simulated evolution using various offsets of the fitness function.

We make multiple conclusions: First, we find that effect of mutations becomes increasingly more detrimental as lower offsets results in more stable proteins. This effect is also observed in real proteins, as recorded in ProTherm [42], although we

observed a 10-fold weaker effect. Second, we demonstrate that a strong coevolution signal arises in our model from direct energetic coupling between neighboring amino acids, and that the coevolution signal is more pronounced with lower fitness function offsets.

The work is still preliminary and exploratory in nature and the analysis is so far based on a single protein.

5.4 Paper IV

One of the key principles that nature follows to generate complex structures is symmetrical self-assembly. Over the past decade, this strategy has also been harnessed by protein engineers and has proven one of the most successful avenues in protein design. In this paper, we review the current literature and discuss the biophysical principles of self-assembly.

5.5 Paper V-VIII

Antibody molecules display a wide-range of sequence and structure diversity because of evolutionary selection pressures that enable reactivity against almost any target antigen. Additionally, antibodies are extremely well-characterized, not only because of their importance to vertebrate life, but also because of their extreme utility in the biopharmaceutical industry. Paper V-VIII takes advantage of this and develop a new method to predict (paper V), design (paper VI and VIII), and study (VII) antibodies.

5.5.1 Paper V

Computational modeling is an essential step in many antibody-engineering workflows. Over the past 20 years the state of the art has relied on homology modeling, expert rules, and in some cases *ab initio* modeling of the most structurally diverse loop of the complementarity determining region (CDR H3). Recent blind benchmarks of antibody modeling strategies showed that existing methods still exhibit modeling inaccuracies in the CDRs, and that models often suffer from stereo-chemical strain. In this paper, we present a new method, AbPredict, for structure prediction that takes advantage of the natural gene segmentation of the antibody structure and the vast number of known and diverse antibody structures.

We begin by extracting the rigid body orientation (RBO) between and the torsion angles of the light and the heavy chain of all antibodies in the Protein Data Bank. Next, we use Rosetta to carry out a simulated annealing Monte Carlo search over antibody backbone fragments corresponding to the natural gene segmentation and RBOs, in order to identify the lowest-energy conformation (Fig. 5.1). The method is

fully automated, uses no expert rules, and does not rely on sequence homology. Although AbPredict is conceptually simple, it performs as well as other methods and is in some cases more accurate. When we investigated the reasons for AbPredict's superior performance we found several cases, where the highest sequence-identity template – which would be the choice of existing methods – exhibits high root-mean-square deviation to the experimental structure; AbPredict instead selects templates of low sequence identity (sometimes <10%) that are more conformationally accurate. In effect, identical or highly similar sequences adopt different conformations, depending on their structural context; conversely, low-homology sequences may converge on very similar conformations, and AbPredict can correctly identify these cases. Furthermore, we found that the models' stereo-chemical properties are uniformly good with the worst model having chemical quality (MolProbity) score of 1.9, compared to >3 for other methods. Due to the high accuracy and low strain of AbPredict's models, they should serve as useful starting points for design and molecular dynamics simulations.

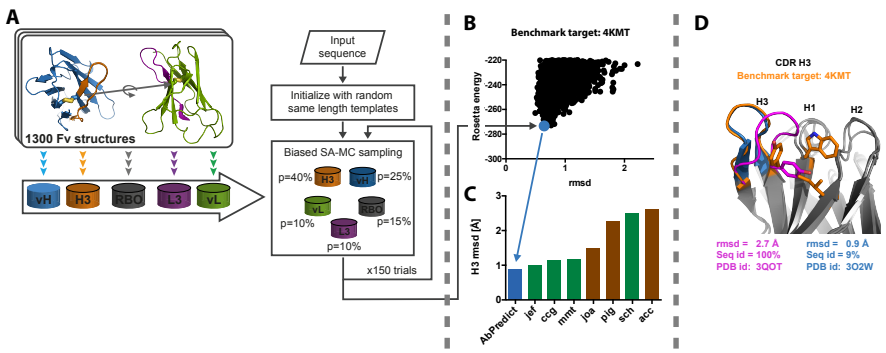


Figure 5.1: *AbPredict* A method for combinatorial backbone modeling of antibodies. (A) The AbPredict algorithm uses a pre-computed database of experimental conformations of antibody structures segmented to reflect genetic V and (D)J recombination. (B) Combinatorial sampling of segments maps out the energy landscape for a target sequence. (C) The lowest energy models are sometimes more accurate than both de novo loop prediction methods (green) and sequence-homology based methods (brown). (D) Often this is due to changes in the structural context. Target structure (orange); highest-identity template (pink); AbPredict template (blue).

5.5.2 Paper VI and VII

Today, new antibody molecules are typically generated by animal immunogenization, which is a tedious process that provides no control over the binding site and often yields proteins too fragile for the reality of industrial expression and clinical use. An outstanding goal has been to build antibodies from first principles [131]. This is what we seek to do in paper VI and VII.

Antibodies bind molecules through loops, which historically have been extremely difficult to accurately design, as the energetic separation between the target loop conformation and alternative conformations is typically small and dominated by energetic contributions that are notoriously difficult to model accurately (water, hydrogen bonds, and backbone torsions). Initially, we relied heavily on the energy function of

Rosetta to design the sequence, but this resulted in antibodies, which barely expressed (Fig. 5.2). Implementing "expert rules", we were able to increase expression-level slightly (cycle 2 and 3). Next, we implemented an evolution-based bias of the energetics following the common assumption of a Boltzmann distributed relationship between amino acid frequencies in alignments and their effect on stability[1]. In the final iteration between computational design and experimental testing, we implemented a new segmentation method, which instead of treating each loop individually, followed the natural gene segmentation of antibodies. This last step increased the expression-level (a proxy for protein stability) to and in some cases beyond the level of an antibody, which had been evolved in vitro for high-expression. The antibody design protocol is described in detail in paper VII.

The antibodies were not only designed to be stable (highly-expressed), they were also designed to bind human insulin or in other cases *Mycobacterium tuberculosis* acyl-carrier protein 2. Out of approximately 100 screened antibodies, we found 3 binding molecules. It was not possible to obtain crystal structures of any of the binding complexes, so it is unclear whether binding was achieved with respect to the intended part of the antigen. However, unbound structures were obtained of the two binding antibodies of the 5^{th} design iteration, and although highly accurate design was achieved throughout most of the protein, the functional binding loops deviated. This was especially true of CDR H3, for which structure-specific evolutionary information is sparse, as it does not have the typical pattern of canonical conformations, which the other CDRs possess.

There can be multiple reasons for the unintended reorganization. First, the protocol did not sample the rigid-body orientation between the light and heavy immunoglobulin domain of the antibody — an important degree of freedom, which we later implemented in paper V. Second, the use of an antibody structure prediction algorithm, might have revealed that conformational rigidity of H3 was not properly encoded. As suggested by the findings of paper V, the same sequence of CDR H3 might adopt widely different conformations depending on its structural environment. Third, problems with the Boltzmann assumption (see Paper I) used to derived the evolutionary energy bias, might have misled the design algorithm. Finally, the necessarily approximate nature of the energy function might have resulted in mistakes.

[1] the research in paper I related to this, was performed after paper VI

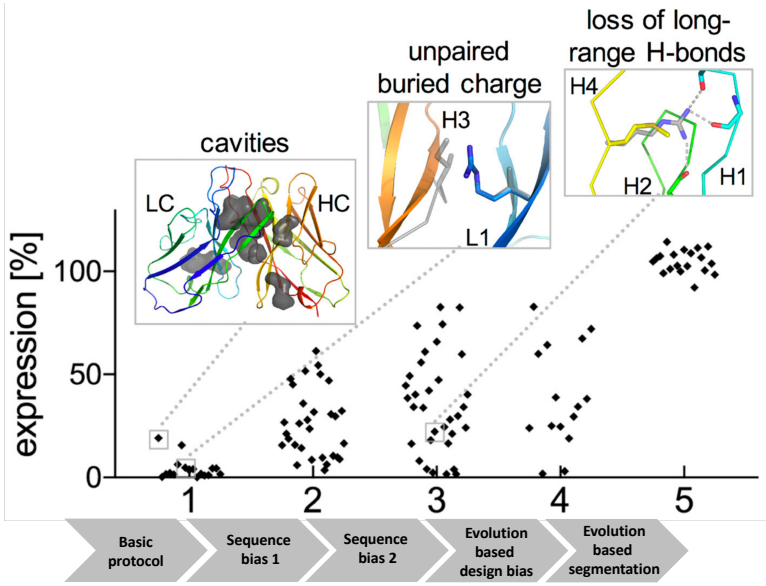


Figure 5.2: Iterative cycle of algorithm development and experimental testing. Expression is normalized relative to an antibody evolved in vitro for high expression levels. Figure adapted from [132].

5.5.3 Paper VIII

The present author's contributions to this work was minimal and only amounted to generating and selecting structures with the algorithm developed in Paper V. Therefore the findings of the paper is only discussed in relation to paper V.

Structural modeling of known antibodies is important and can provide biochemical insight necessary for further engineering. In paper VII, an antibody targeting a tumor-associated carbohydrate antigen is studied. One of the salient qualities of the AbPredict algorithm of paper V is the high chemical quality of the resulting structures, which should make them particularly useful for molecular dynamics simulations. In paper VII, an AbPredict antibody model is subjected to a 500 ns molecular dynamics simulation, and the structural fluctuation are found to be similar to that of an antibody crystal structure. Given that only a single model was simulated, this only serves as a limited demonstration of the utility of AbPredict.

5.6 Paper IX

This paper concerns data-driven modeling of a protein structure. It is unconnected from the evolutionary theme many of the other papers have followed..

Some proteins contain regions so flexible and disordered that there is no reasonable hope that they could ever be characterized as a single conformation or for that sake crystallized. Calreticulin is one such protein, which in an otherwise well-structure core, has a long proline-rich region (P-domain) and a C-terminal (C-domain) tail, which both appear largely disordered and together constitute more than half of the 400 residues long protein. Calreticulin is crucial for chaperoning glycoproteins and maintaining intracellular Ca^{2+} homeostasis. We sought to understand how the changes in calcium concentration could affect the accessibility of the carbohydrate-binding region of calreticulin (the lectin site).

To gain information about the position of the P and the C domain, chemical cross-linking data was gathered. Specifically, a lysine-reactive cross-linker was randomly reacted with lysine, whereafter the protein was digested with proteases, and the resulting fragments analyzed with mass spectrometry. Fragments that were bound with the cross-linker revealed lysine situated at most 21.4 Å apart (the length of the cross-linker as well as two extended lysines). As such cross-links provide quite inaccurate estimates of contact lengths. Using a soft-threshold classification for when cross-links were broken, we generated and analyzed a structural ensemble of calreticulin with Rosetta.

We found that the resulting models appeared to deviate depending on whether cross-links were collected in the presence or absence of calcium. In the presence of calcium, the resulting structural ensemble of the P-loop tended to cover the lectin-binding site. The findings suggest that the role of calcium binding in calreticulin is not only related to calcium storage but might also be involved in regulating the chaperoning activity.

References

[1] Dobzhansky, T. Nothing in Biology Makes sense except in the Light of Evolution. *Am. Biol. Teach.* **1973**, *35*, 125–129.

[2] Dodd, M. S.; Papineau, D.; Grenne, T.; Slack, J. F.; Rittner, M.; Pirajno, F.; O'Neil, J.; Little, C. T. Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* **2017**, *543*, 60–64.

[3] Box, G. *Robustness Stat.*; Academic Press Inc., 1979; pp 201–236.

[4] Box, G. Science and Statistics. *J. Am. Stat. Assoc.* **1976**, *71*, 791–799.

[5] Darwin, C. *Orig. Species by Means Nat. Sel. or Preserv. Favoured Races Struggl. Life.*; John Murray, 1859; p 386.

[6] Darwin, C.; Wallace, A. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *J. Proc. Linn. Soc. London* **1858**, *3*, 45–62.

[7] Kimura, M. Some Problems of Stochastic Processes in Genetics. *Ann. Math. Stat.* **1957**, *28*, 882–901.

[8] Kimura, M. Evolutionary Rate at the Molecular Level. *Nature* **1968**, *217*, 624–626.

[9] Watson, J.; Crick, F. Moelcular structure of nucleic acids. *Nature* **1953**, *171*, 256–257.

[10] Hardy, G. Mendelian proportions in a mixed population. *Science* **1908**, *XXVIII*, 49–50.

[11] May, R. Uses and Abuses of Mathematics in Biology. *Science* **2004**, *303*, 790–793.

[12] Lee, H.; Popodi, E.; Tang, H.; Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *PNAS* **2012**, *109*, 2774–2783.

[13] Reid, T. M.; Loebt, L. A.; Gottstein, J. Tandem double CC-> TT mutations are produced by reactive oxygen species. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 3904–3907.

[14] Harris, K.; Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res.* **2014**, *24*, 1445–1454.

[15] Orr, H. A. Fitness and its role in evolutionary genetics. *Nat. Rev. Genet.* **2009**, *10*, 531–539.

[16] Nowak, M.; Tarnita, C.; Wilson, E. The evolution of eusociality. *Nature* **2010**, *446*, 1057–1062.

[17] Nowak, M. *Belkn. Press Harvard Univ. Press*; 2006; Vol. 1.

[18] Kimura, M. On the probability of fixation of mutant gennes in a population. *Genetics* **1962**, *47*, 713–719.

[19] Sella, G.; Hirsh, A. The application of statistical physics to evolutionary biology. *PNAS* **2005**, *102*, 12690–12693.

[20] Cvijović, I.; Good, B. H.; Jerison, E. R.; Desai, M. M. Fate of a mutation in a fluctuating environment. *Proc. Natl. Acad. Sci.* **2015**, *112*, E5021–E5028.

[21] Hauser, O. P.; Traulsen, A.; Nowak, M. A. Heterogeneity in background fitness acts as a suppressor of selection. *J. Theor. Biol.* **2014**, *343*, 178–185.

[22] Whitlock, M. C.; Gomulkiewicz, R. Probability of fixation in a heterogeneous environment. *Genetics* **2005**, *171*, 1407–1417.

[23] Whitlock, M. C. Fixation probability and time in subdivided populations. *Genetics* **2003**, *164*, 767–779.

[24] Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225*, 563–564.

[25] Taylor, C.; Iwasa, Y.; Nowak, M. A. A symmetry of fixation times in evoultionary dynamics. *J. Theor. Biol.* **2006**, *243*, 245–251.

[26] Iwasa, Y. Free Fitness that Always Increases in Evolution. *J Theor Biol* **1988**, *135*, 265–281.

[27] Felsenstein, J. *Inferring Phylogenies*; Oxford University Press, 2003; p 580.

[28] Mayrose, I.; Graur, D.; Ben-Tal, N.; Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Mol. Biol. Evol.* **2004**, *21*, 1781–1791.

[29] Lunter, G.; Miklós, I.; Drummond, A.; Jensen, J. L.; Hein, J. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **2005**, *6*, 1–10.

[30] Katoh, K.; Misawa, K.; Kuma, K.-i.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066.

[31] Brown, A. *J.D. Bernal: The Sage of Science*; Oxford University Press, 2005; p 343.

[32] Hershberg, R.; Petrov, D. A. Selection on Codon Bias. *Annu. Rev. Genet.* **2008**, *42*, 287–299.

[33] Plotkin, J. B.; Kudla, G. Synonymous but not the same: The causes and consequences of codon bias. *Nat. Rev. Genet.* **2011**, *12*, 32–42.

[34] Allan Drummond, D.; Wilke, C. O. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* **2009**, *10*, 715–724.

[35] Echave, J.; Wilke, C. O. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annu. Rev. Biophys.* **2017**, *46*, 85–103.

[36] Fleishman, S. J.; Khare, S. D.; Koga, N.; Baker, D. Restricted sidechain plasticity in the structures of native proteins and complexes. *Protein Sci.* **2011**, *20*, 753–757.

[37] Karanicolas, J.; Corn, J. E.; Chen, I.; Joachimiak, L. a.; Dym, O.; Peck, S. H.; Albeck, S.; Unger, T.; Hu, W.; Liu, G.; Delbecq, S.; Montelione, G. T.; Spiegel, C. P.; Liu, D. R.; Baker, D. A de novo protein binding pair by computational design and directed evolution. *Mol. Cell* **2011**, *42*, 250–60.

[38] Hu, X.; Wang, H.; Ke, H.; Kuhlman, B. High-resolution design of a protein loop. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 17668–73.

[39] Fleishman, S. J.; Baker, D. Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution. *Cell* **2012**, *149*, 262–273.

[40] Drummond, D. A.; Bloom, J. D.; Adami, C.; Wilke, C. O.; Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci.* **2005**, *102*, 14338–14343.

[41] Wylie, C. S.; Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness in viruses. *Proc. Natl. Acad. Sci.* **2011**, *108*, 9916–9921.

[42] Serohijos, A. W. R.; Rimas, Z.; Shakhnovich, E. I. Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. *Cell Rep.* **2012**, *2*, 249–256.

[43] Kumar, M. D. S. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* **2006**, *34*, D204–D206.

[44] Tian, P.; Best, R. B. How Many Protein Sequences Fold to a Given Structure? A Coevolutionary Analysis. *Biophys. J.* **2017**, *113*, 1719–1730.

[45] Yang, J.-R.; Liao, B.-Y.; Zhuang, S.-M.; Zhang, J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl. Acad. Sci.* **2012**, *109*, E831–E840.

[46] Levy, E. D.; De, S.; Teichmann, S. A. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc. Natl. Acad. Sci.* **2012**, *109*, 20461–20466.

[47] Richardson, J. S.; Richardson, D. C.; Tweedy, N. B.; Gernert, K. M.; Quinn, T. P.; Hecht, M. H.; Erickson, B. W.; Yan, Y.; Mcclain, R. D.; Donlan, M. E.; Surles, M. C. Looking at proteins: representations, folding, packing, and design. *Biophys. J* **1992**, *63*, 1186–1209.

[48] Richardson, J. S.; Richardson, D. C. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 2754–9.

[49] Dekel, E.; Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **2005**, *436*, 588–592.

[50] Akashi, H.; Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 3695–3700.

[51] Shoichet, B. K.; Baase, W. A.; Kuroki, R.; Matthews, B. W. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci.* **1995**, *92*, 452–456.

[52] Beadle, B. M.; Shoichet, B. K. Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.* **2002**, *321*, 285–296.

[53] Elcock, A. H. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* **2001**, *312*, 885–896.

[54] Pei, J.; Dokholyan, N. V.; Shakhnovich, E. I.; Grishin, N. V. Using protein design for homology detection and active site searches. *Proc. Natl. Acad. Sci.* **2003**, *100*, 11361–11366.

[55] Chelliah, V.; Chen, L.; Blundell, T. L.; Lovell, S. C. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **2004**, *342*, 1487–1504.

[56] Cheng, G.; Qian, B.; Samudrala, R.; Baker, D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.* **2005**, *33*, 5861–5867.

[57] Ghosh, K.; Dill, K. Cellular proteomes have broad distributions of protein stability. *Biophys. J.* **2010**, *99*, 3996–4002.

[58] DePristo, M. a.; Weinreich, D. M.; Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **2005**, *6*, 678–87.

[59] Taverna, D. M.; Goldstein, R. A. Why are proteins marginally stable? *Proteins Struct. Funct. Genet.* **2002**, *46*, 105–109.

[60] Williams, P. D.; Pollock, D. D.; Goldstein, R. a. Functionality and the evolution of marginal stability in proteins: inferences from lattice simulations. *Evol. Bioinform. Online* **2006**, *2*, 91–101.

[61] Goldstein, R. A. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins Struct. Funct. Bioinforma.* **2011**, *79*, 1396–1407.

[62] Hietpas, R. T.; Jensen, J. D.; Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci.* **2011**, *108*, 7896–7901.

[63] Sanjuan, R.; Moya, A.; Elena, S. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. **2004**,

[64] Firnberg, E.; Labonte, J. W.; Gray, J. J.; Ostermeier, M. A comprehensive, high-resolution map of a Gene's fitness landscape. *Mol. Biol. Evol.* **2014**, *31*, 1581–1592.

[65] Tokuriki, N.; Stricher, F.; Schymkowitz, J.; Serrano, L.; Tawfik, D. S. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *J. Mol. Biol.* **2007**, *369*, 1318–1332.

[66] Balakrishnan, S.; Kamisetty, H.; Carbonell, J. G.; Lee, S. I.; Langmead, C. J. Learning generative models for protein fold families. *Proteins Struct. Funct. Bioinforma.* **2011**, *79*, 1061–1078.

[67] Ovchinnikov, S.; Kamisetty, H.; Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **2014**, *2014*, 1–21.

[68] Talavera, D.; Lovell, S. C.; Whelan, S. Covariation is a poor measure of molecular coevolution. *Mol. Biol. Evol.* **2015**, *32*, 2456–2468.

[69] Poelwijk, F. J.; T□nase-Nicola, S.; Kiviet, D. J.; Tans, S. J. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J. Theor. Biol.* **2011**, *272*, 141–144.

[70] Pollock, D. D.; Thiltgen, G.; Goldstein, R. a. Amino acid coevolution induces an evolutionary Stokes shift. *Proc. Natl. Acad. Sci.* **2012**, *109*, E1352–E1359.

[71] Shah, P.; McCandlish, D. M.; Plotkin, J. B. Contingency and entrenchment in protein evolution under purifying selection. *Proc. Natl. Acad. Sci.* **2015**, *112*, E3226–E3235.

[72] Naumenko, S. A.; Kondrashov, A. S.; Bazykin, G. A. Fitness conferred by replaced amino acids declines with time. *Biol. Lett.* **2012**, *8*, 825–828.

[73] Goldstein, R. A.; Pollard, S. T.; Shah, S. D.; Pollock, D. D. Nonadaptive amino acid convergence rates decrease over time. *Mol. Biol. Evol.* **2015**, *32*, 1373–1381.

[74] Perutz, M. F.; Kendrew, J. C.; Watson, H. C. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **1965**, *13*, 669–678.

[75] Goldman, N.; Thorne, J. L.; Jones, D. T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **1998**, *149*, 445–458.

[76] Worth, C. L.; Gong, S.; Blundell, T. L. Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 709–720.

[77] Franzosa, E. A.; Xia, Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **2009**, *26*, 2387–2395.

[78] Echave, J.; Spielman, S. J.; Wilke, C. O. Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **2016**, *17*, 109–121.

[79] Jiang, Q.; Teufel, A.; Jackson, E.; Wilke, C. Beyond Thermodynamic Constraints: Evolutionary Sampling Generates Realistic Protein Sequence Variation. *Genetics* **2018**, *208*, 1387–1395.

[80] Echave, J.; Jackson, E. L.; Wilke, C. O. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Phys. Biol.* **2015**, *12*.

[81] Bloom, J. D.; Glassman, M. J. Inferring Stabilizing Mutations from Protein Phylogenies: Application to Influenza Hemagglutinin. *PLoS Comput. Biol.* **2009**, *5*, e1000349.

[82] Dessailly, B. H.; Lensink, M. F.; Wodak, S. J. Relating destabilizing regions to known functional sites in proteins. *BMC Bioinformatics* **2007**, *8*, 1–22.

[83] Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M. Analysis of Catalytic Residues in Enzyme Active Sites. *J. Mol. Biol.* **2002**, *324*, 105–121.

[84] Pupko, T.; Bell, R. E.; Mayrose, I.; Glaser, F.; Ben-Tal, N. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **2002**, *18*, 71–77.

[85] Celniker, G.; Nimrod, G.; Ashkenazy, H.; Glaser, F.; Martz, E.; Mayrose, I.; Pupko, T.; Ben-Tal, N. ConSurf: Using evolutionary data to raise testable hypotheses about protein function. *Isr. J. Chem.* **2013**, *53*, 199–206.

[86] Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **1992**, *89*, 10915–10919.

[87] Kosiol, C.; Holmes, I.; Goldman, N. An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* **2007**, *24*, 1464–1479.

[88] Whelan, S.; Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **2001**, *18*, 691–699.

[89] Le, S. Q.; Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **2008**, *25*, 1307–1320.

[90] Atchley, W. R.; Zhao, J.; Fernandes, A. D.; Druke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci.* **2005**, *102*, 6395–6400.

[91] Tomii, K.; Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* **1996**, *9*, 27–36.

[92] Venkatarajan, M.; Braun, W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *J. Mol. Model.* **2001**, *7*, 445–453.

[93] Overington, J.; Donnelly, D.; Johnson, M.; Sali, A.; Blundell, T. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci.* **1992**, *1*, 216–226.

[94] Le, S. Q.; Dang, C. C.; Gascuel, O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.* **2012**, *29*, 2921–2936.

[95] Dellus-Gur, E.; Toth-Petroczy, A.; Elias, M.; Tawfik, D. S. What makes a protein fold amenable to functional innovation? fold polarity and stability trade-offs. *J. Mol. Biol.* **2013**, *425*, 2609–2621.

[96] Hirano, M.; Das, S.; Guo, P.; Cooper, M. D. *Adv. Immunol.*, 1st ed.; Elsevier inc., 2011; Vol. 109; pp 125–157.

[97] Glanville, J.; Zhai, W.; Berka, J.; Telman, D.; Huerta, G.; Mehta, G. R.; Ni, I.; Mei, L.; Sundar, P. D.; Day, G. M. R.; Cox, D.; Rajpal, A.; Pons, J. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci.* **2009**, *106*, 20216–20221.

[98] Al-Lazikani, B.; Lesk, A.; Chothia, C. Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* **1997**, *273*, 927–48.

[99] Owen, J.; Punt, J.; Strandford, S. *Immunology*; W. H.. Freeman, 2013; p 574.

[100] James, L. C.; Roversi, P.; Tawfik, D. S. Antibody multispecificity mediated by conformational diversity. *Science* **2003**, *299*, 1362–7.

[101] Tokuriki, N.; Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **2009**, *19*, 596–604.

[102] Lapidoth, G. Computational Design of Protein Function Using Modular Backbone Assembly, http://docs.wixstatic.com/ugd/11c012_e83bfaffbe144b1989c9b120c0117cf7.pdf. 2018.

[103] Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13*, 3031–3048.

[104] Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **1973**, *181*, 223–230.

[105] Govindarajan, S.; Goldstein, R. A. On the thermodynamic hypothesis of protein folding. *Proc. Natl. Acad. Sci.* **1998**, *95*, 5545–5549.

[106] Park, H.; Bradley, P.; Greisen, P.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; Dimaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **2016**, *12*, 6201–6212.

[107] Sippl, M. J. Calculation of conformational ensembles from potentials of mena force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859–883.

[108] Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209–25.

[109] Hamelryck, T.; Borg, M.; Paluszewski, M.; Paulsen, J.; Frellsen, J.; Andreetta, C.; Boomsma, W.; Bottaro, S.; Ferkinghoff-Borg, J. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS One* **2010**, *5*, e13714.

[110] Kuhlman, B.; Dantas, G.; Ireton, G.; Varani, G.; Stoddard, B.; Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364–8.

[111] Steipe, B.; Schiller, B.; Plückthun, A.; Steinbacher, S. Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* **1994**, *240*, 188–192.

[112] Huang, P. S.; Boyken, S. E.; Baker, D. The coming of age of de novo protein design. *Nature* **2016**, *537*, 320–327.

[113] Marcos, E. et al. Principles for designing proteins with cavities formed by curved β sheets. *Science* **2017**, *355*, 201–206.

[114] Dou, J. et al. De novo design of a fluorescence-activating β-barrel. *Nature* **2018**, *561*, 485–491.

[115] Norn, C. H.; André, I. Computational design of protein self-assembly. *Curr. Opin. Struct. Biol.* **2016**, *39*, 39–45.

[116] Havranek, J.; Harbury, P. Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **2003**, *10*, 45–52.

[117] Leaver-Fay, A.; Froning, K. J.; Atwell, S.; Aldaz, H.; Pustilnik, A.; Lu, F.; Huang, F.; Yuan, R.; Hassanali, S.; Chamberlain, A. K.; Fitchett, J. R.; Demarest, S. J.; Kuhlman, B. Computationally Designed Bispecific Antibodies using Negative State Repertoires. *Structure* **2016**, *24*, 641–651.

[118] Netzer, R.; Listov, D.; Lipsh, R.; Dym, O.; Albeck, S.; Knop, O.; Kleanthous, C.; Fleishman, S. J. Ultrahigh specificity in a network of computationally designed protein-interaction pairs. *Nat. Commun.* **2018**, *9*, 5286.

[119] Fleishman, S.; Corn, J.; Strauch, E.; Whitehead, T.; Karanicolas, J.; Baker, D. Hotspot-centric de novo design of protein binders. *J. Mol. Biol.* **2011**, *413*, 1047–62.

[120] Fleishman, S. J.; Whitehead, T. a.; Ekiert, D. C.; Dreyfus, C.; Corn, J. E.; Strauch, E.-M.; Wilson, I. a.; Baker, D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **2011**, *332*, 816–21.

[121] Boyken, S. E.; Chen, Z.; Groves, B.; Langan, R. A.; Oberdorfer, G.; Ford, A.; Gilmore, J. M.; Xu, C.; DiMaio, F.; Pereira, J. H.; Sankaran, B.; Seelig, G.; Zwart, P. H.; Baker, D. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **2016**, *352*, 680–687.

[122] Komor, R. S.; Romero, P. A.; Xie, C. B.; Arnold, F. H. Highly thermostable fungal cellobiohydrolase i (Cel7A) engineered using predictive methods. *Protein Eng. Des. Sel.* **2012**, *25*, 827–833.

[123] Goldenzweig, A. et al. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **2016**, *63*, 337–346.

[124] Khersonsky, O.; Lipsh, R.; Avizemer, Z.; Ashani, Y.; Goldsmith, M.; Leader, H.; Dym, O.; Rogotner, S.; Trudeau, D. L.; Prilusky, J.; Amengual-Rigo, P.; Guallar, V.; Tawfik, D. S.; Fleishman, S. J. Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* **2018**, *72*, 178–186.e5.

[125] Dunbrack, R.; Karplus, M. Backbone-dependent rotamer library for proteins. *J.* **1993**, *230*, 543–573.

[126] Shapovalov, M. V.; Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19*, 844–858.

[127] Kuhlman, B.; Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci.* **2000**, *97*, 10383–10388.

[128] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[129] Abagyan, R.; Kuznetsov, D.; Totrov, M. ICM - New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from. *J. Comput. Chem.* **1994**, *15*, 488–506.

[130] Conway, P.; Tyka, M. D.; DiMaio, F.; Konerding, D. E.; Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **2014**, *23*, 47–55.

[131] Winter, G.; Milstein, C. Man-made antibodies. *Nature* **1991**, *349*, 293–299.

[132] Baran, D.; Pszolla, M. G.; Lapidoth, G. D.; Norn, C.; Dym, O.; Unger, T.; Albeck, S.; Tyka, M. D.; Fleishman, S. J. Principles for computational design of binding antibodies. *Proc. Natl. Acad. Sci.* **2017**, *114*, 10900–10905.

Chapter 6

Scientific publications

Author contributions

Paper I: A biophysical model of protein evolution relates amino acid frequencies and protein stability

I took part in conceiving and designing the research, analysis, and deriving the mathematical framework. I wrote the initial draft of the manuscript and took part in revisions.

Paper II: A thermodynamic model of protein structure evolution explains amino acid rate matrices and predicts functional sites

I took part in conceiving and designing the project, derivation of the mathematical framework, and data generation and analysis. I wrote the initial draft of the manuscript and took part in revisions.

Paper III: Properties of all-atom simulations of protein evolution

I implemented the evolutionary model, the tree population algorithm, and several of the methods for analyzing the trajectories. I took part in analyzing the data.

Paper IV: Computational design of protein self-assembly

I took part in reviewing the literature and writing the paper.

Paper v: High-accuracy modeling of antibody structures by a search for minimum-energy recombination of backbone fragments

I took part in conceiving and designing the research and in developing the method. I generated the data. I took part in data analysis and in writing the paper.

Paper vi: Principles for computational design of binding antibodies

I contributed to algorithm development, developed tools for rapid structure visualization and analysis, and designed and experimentally screened a subset of antibodies.

Paper vii: A combined computational-experimental approach to define the structural origin of antibody recognition of sialyl-Tn, a tumor-associated carbohydrate antigen

I contributed to algorithm development, data analysis, and to the idea of V-(D)J segmentation.

Paper viii: AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences

I generated structures with AbPredict and took part in the analysis of ditto.

Paper ix: Mapping the Ca2+ induced structural change in calreticulin

I designed and carried out the structural analysis and wrote parts of the manuscript.

Printed by Media-Tryck, Lund 2019 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculty of Science
Department of Biochemistry and Structural Biology

ISBN 978-91-7422-618-8