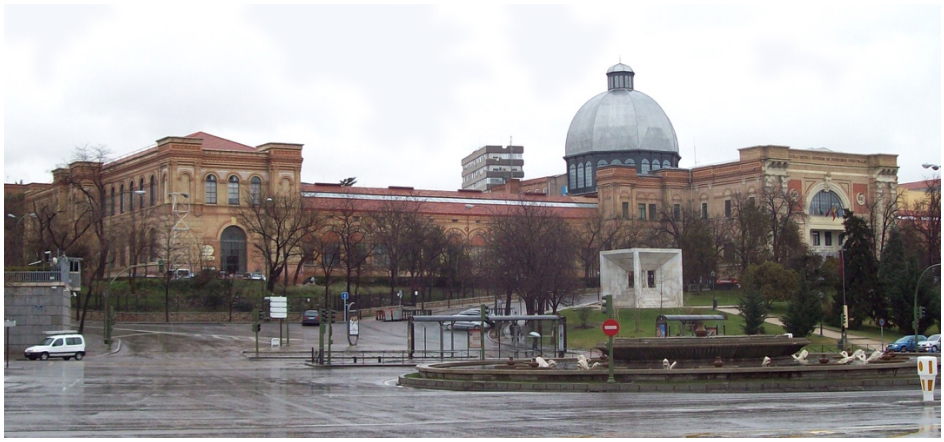
Symposium "Evolutionary intelligence: How coevolution opened the way to the AlphaFold revolution"



Venue: National Museum of Natural Sciences-CSIC

Madrid (Spain)

27-28th August 2025

Program &

Book of Abstracts

Sponsored by:

JOURNAL OF MOLECULAR EVOLUTION

CSIC HUB BIOLOGÍA COMPUTACIONAL Y BIOINFORMÁTICA (BCB) CSIC

mncn 250 1771 2021 museo nacional de ciencias naturales

Program

Venue: Salón de actos

Wednesday, August 27th

15:45-16:00 Opening (D. Juan, R. Zardoya and A. Rojas)

16:00-17:30 Section I (3T)

- **Alfonso Valencia**, Protein sequence-structure-function relationships: from coevolution to deep learning
- **José Onuchic (R)**, Energy Landscapes, Order and Disorder, and Protein Sequence Coevolution: From Proteins to Chromosome Structure
- **Chris Sander (R)**, 3Dseq: From experimental evolution to 3D structures using EVfold

17:30-18:00 Coffee break

18:00-19:30 Section 2 (3T)

- **David Juan,** Across layers and species: comparative omics to infer functional interactions
- **David Jones (R),** Covariation to AlphaFold: A Compressed History

Thursday, August 28th

9:00-10:30 Section 3 (3T)

- **Erik Aurell (R)**, Inferring fitness from large genomic data sets
- **Laura Orellana**, Protein Dynamics, disease mutations and coevolution - case studies in cancer
- **Andrea Pagnani (R)**, Inferring protein landscapes from co-evolutionary and selection data

10:30-11:00 Coffee break

11:00-12:30 Section 4 (3T)

- **Christine Orengo**, AlphaFold massively expands CATH superfamilies giving insights into protein evolution and functional mechanisms
- **Jessica Siltberg-Liberles (R)**, Using AlphaFold2 for studying protein family evolution
- **Ana M. Rojas,** Addressing protein function and similarity via transformers

12:30-14:00 lunch

14:00:15:30 Section 5 (3T)

- **David Talavera,** Covariation and coevolution: why the difference matters (to me)
- **Doug Barrick,** Covariant residue pairs reveals are important for enzyme activity but not stability

- **Martin Weigt (R)**, Protein Evolution in Sequence Landscapes - From Data to Models and Back

15:30-16:00 Coffee break

16:00-17:30 Section 6 (3T)

- **Modesto Orozco,** Coevolutionary contacts guiding the determination of alternative folds in proteins:
- **Simona Cocco (R)**, Generative models learned on sequence data to forecast SARSCov2 viral evolution and antibody resilience

17:30-19:00 Round table (moderators Valencia & Orengo)

Abstracts

(ordered by submission date, to be ordered after approving the groupings)

1. Protein Sequence-Structure-Function Relationships: From Coevolution to Deep Learning

Alfonso Valencia,

ICREA and Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain.

For years, Multiple Sequence Alignments (MSAs) have been the primary source of evolutionary information, enabling the development of methods to detect coevolution between protein positions at various levels. However, these approaches were never sufficiently effective for guiding protein structure prediction or reliably distinguishing binding/functional sites from other conserved regions. The advent of AI for structure prediction and for the generation of vast collections of "viable" protein sequences has revolutionized this field, enabling the simultaneous analysis of conservation at the sequence and structural level (conservation of biophysical properties).

Leveraging these advances, we have begun large-scale analyses of protein families to investigate how proteins accommodate structural features, functional sites, and foldability and how they adapt along time. In this talk, I will present our current perspectives on how binding and functional sites evolve within the constraints of protein foldability and what could be the consequences for protein design and functional diversification.

2. Energy Landscapes, Order and Disorder, and Protein Sequence Coevolution: From Proteins to Chromosome Structure

José Nelson Onuchic*

Center for Theoretical Biological Physics and Departments of Physics and Astronomy, Chemistry, and Biosciences. Rice University, Houston, Texas, USA

Energy landscape theory has been a powerful approach to study protein folding dynamics and function. The discovery that an accurate estimate of the joint probability distribution of amino acid occupancies in protein families provides insights about residue-residue coevolution and concrete details about protein folding landscapes has also advanced structural biophysics. Our realization that the collection of couplings and local fields as parameters of such distribution is inherently connected with the thermodynamics of sequence selection towards folding and function demonstrates the importance of coevolutionary methods to understand stability and function of biomolecules. The synergy between structure based models and coevolutionary information has spearheaded the field of structure prediction, including protein and RNA, as well as accelerating the discovery of functional structural states and the prediction of protein complexes. Coevolution signals can also be used to create protein recognition metrics, which led to successful experimental efforts, and the uncovering of novel molecular interactions. This idea has opened the door to encode recognition in protein pairs. Recently this approach has been used to predict extremely large protein assemblies consisting of structural maintenance of chromosomes (SMC) and kleisin subunits which are essential for the process of

chromosome segregation across all domains of life. While limited structural data exist for the proteins that comprise the (SMC)–kleisin complex, using an integrative approach combining both crystallographic data and coevolutionary information, we have predicted an atomic-scale structure of the whole condensing complex in prokaryotes. These ideas from co-evolution can also be utilized in genome folding and function. The energy landscape of the model was derived by using the maximum entropy principle and relies on two experimentally derived inputs: a classification of loci into chromatin types and a catalog of the positions of chromatin loops. This model was generalized by utilizing a neural network to infer these chromatin types using epigenetic marks present at a locus, as assayed by ChIP-Seq. The ensemble of structures resulting from these simulations completely agree with HI-C data and exhibits unknotted chromosomes, phase separation of chromatin types, and a tendency for open chromatin to lie at the periphery of chromosome territories.

* supported by the NSF and the Welch Foundation

3. 3Dseq: From experimental evolution to 3D structures using EVfold

Chris Sander

Systems Biology, Harvard Medical School, Boston, MA, USA

We have explored a fourth experimental method of protein structure determination using evolution experiments in the laboratory, called 3Dseq. For a protein of interest, the experiments involve sequence variation in a library of millions of sequences to which functional or structural assays are applied that select a performant subset of sequences. After one or more rounds of generation of variant sequence libraries and functional selection, the resultant sequence libraries of many thousands of sequences provide rich information about constrained residue-residue interactions. These interaction constraints can be used to compute correct protein 3D folds that are accurate to within a few Angstrom of positional variation of protein coordinates compared to static crystal structures. The interaction constraints also identify functionally important interactions that are informative for quantitative evolutionary biology, protein design and for the development of drug therapies. The 3Dseq method of protein structure determination complements the classic methods of X-ray crystallography, NMR spectroscopy and cryo-EM electron microscopy and provides a tool for exploring molecular evolution extrapolated to the future. Publications: bit.ly/3DseqOpen, journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766

4. Across layers and species: comparative studies to infer functional interactions

David Juan

Functional and Comparative Multiomics, National Center for Biotechnology (CNB-CSIC), Madrid, Spain

Biological systems are organized into complex networks of coordinated interactions among molecular components, such as proteins, genes, and regulatory elements, that together shape phenotypic outcomes. These interactions define functional and structural scaffolds where evolutionary conservation and innovation take place and are characterized by spatiotemporal, biochemical, and evolutionary constraints. Patterns of covariation within and between species

across diverse biological features, including protein sequences, genome occupancy, regulatory activity, and species traits, can reveal underlying functional dependencies shaped by these constraints.

In this presentation, starting from coevolution, I will revisit several studies leveraging comparative data to uncover and characterize functional dependencies. These include our efforts to disentangle gene regulatory interactions, characterize regulatory evolution, and infer genome–phenome associations underlying traits such as cancer prevalence in primates. Through these case studies, I will illustrate how covariation-based frameworks, when integrated with evolutionary analysis, can uncover system-level properties and functional relationships often inaccessible through single-target or single-species analyses.

5. Covariation to AlphaFold: A Compressed History

David Jones

Institute of Structural and Molecular Biology, University College London, London WC1E 6BT, UK
Department of Computer Science, University College London, London WC1E 6BT, UK

This talk will look back at the journey from early residue–residue covariation to today's deep learning models like AlphaFold, told from a personal point of view. Along the way, we'll revisit some of the methods that made key leaps possible — from statistical contact prediction to neural networks — and see how, at their core, they all relied on the same basic trick: squeezing complex sequence data down to the bits that really matter for structure. Rather than a deep technical dive, this will be more of a guided tour of the ideas, turning points, and surprises that shaped the field, and how the theme of "compression" quietly connects them all.

6. Inferring fitness from large genomic data sets

Erik Aurell

AlbaNova University Centre, KTH-Royal Institute of Technology, SE-106 91 Stockholm, Sweden

Throughout the course of the SARS-CoV-2 pandemic, genetic variation contributed to the spread and persistence of the virus. For example, various mutations have allowed SARS-CoV-2 to escape antibody neutralization or to bind more strongly to the receptors that it uses to enter human cells. In the course of the pandemic more viral genomes (tens of millions) were produced than in any previous pandemic, which can then be used in ways that could not be envisaged in the past.

I will discuss two methods to infer genetic fitness (additive and epistatic fitness) from such data sets, and compare them to simulations where underlying parameters are known. I will then also discuss the implications for future pandemics, and what we could (plausibly) do with similar types of data in other settings.

The talk is based on joint work with Hongli Zeng and John Barton, partly published in Physical Biology 22:016003 (2024), and partly in preparation.

7. Protein Dynamics, disease mutations and coevolution - case studies in cancer

Laura Orellana

Protein Dynamics and Mutation Lab, Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden

Proteins conform the ultimate machinery of Life, executing all processes that sustain living organisms – from complex metabolic pathways to neurotransmission. Far from being static, at physiological temperatures, proteins vibrate and cycle between different states or conformers, sensing external signals: in the same way that primary sequences fold into 3D-sructures, each shape encodes intrinsic functional motions of such relevance for Life that are conserved from bacteria to humans [1]. Nevertheless, although protein conformational mechanisms are key to understand the link between structure and function, they are often elusive for both experiments and simulations. To overcome these limitations and explore them in-depth, our research integrates coarse-grained and atomistic simulations [2-4] to screen pathogenic mutations in cancer and mendelian diseases, which can reveal unexpected mechanisms validated up to the in vitro and in vivo level in mice [5]. Here we present preliminary research on selected examples where these multiscale approaches, together with evolutionary analysis, suggest potential connections between key residues mutated in cancer and co-evolved residue pairs

Keywords: Structural and Computational Biophysics & Bioinformatics, Protein Dynamics, Coarse-Grained Modelling, Elastic Network Models, Mutations

References:

[1] L. Orellana, Are Protein Shape-Encoded Lowest-Frequency Motions a Key Phenotype Selected by Evolution? Applied Sciences, 13(11): 6756, 2023
[2] L. Orellana, Large-scale conformational changes and protein function: breaking the in silico barrier, Frontiers in Molecular Biosciences, 6:117, 2019
[3] Orellana L, Yoluk O, Carrillo O, Orozco M, Lindahl E. (2016). Prediction and validation of protein intermediate states from structurally rich ensembles and coarse-grained simulations. Nat. Comms; doi:10.1038/ncomms12575
[4] Scaramozzino D., Lee B.H. and Orellana L. (2025), Breaking the size limit: efficient sampling of large-scale transition pathways and intermediate conformations in sub-mesoscopic protein complexes. Nat. Comms. (awaiting resubmission); doi: 10.21203/rs.3.rs-6504036/v1
[5] L. Orellana, Thorne AH, et al Oncogenic mutations at the EGFR ectodomain structurally converge to remove a steric hindrance on a kinase-coupled cryptic epitope. PNAS, 116 (20): 10009-10018, 2019

8. Inferring protein landscapes from co-evolutionary and selection data

Andrea Pagnani

DISAT, Politecnico di Torino, Torino, Italy
Italian Institute for Genomic Medicine, Candiolo, Italy
INFN, Sezione di Torino, Torino, Italy

In the last few years, the development of increasingly accurate high-throughput biochemical assays with massively parallel sequencing techniques has established large-scale genetic screening as a fundamental tool for the investigation of the relationship between evolution, fitness, and other critical biological concepts that are behind experimental research. I will describe the two main types of screening experiments - deep mutational scanning and directed evolution - and the different inference frameworks we developed to analyze them.

9 AlphaFold massively expands CATH superfamilies giving insights into protein evolution and functional mechanisms

Christine Orengo

Institute of Structural and Molecular Biology, University College London, London, UK

The recent development of the AlphaFold2 method by DeepMind, has led to a massive expansion in high quality protein structure data. Our group have developed computational protocols (Chainsaw, CATHe, CATH-AlphaFlow) to classify these structural data into evolutionary families. In collaboration with the group of David Jones, also at UCL, we have classified >200 million predicted structures in the AlphaFold database into evolutionary families. This information is available in a new resource – TED – The Encyclopaedia of Domains and also via the AlphaFold Database at the EBI. My group have also developed methods for subclassifying proteins in evolutionary families into functional families. In the talk I will present some insights from TED and describe how we are using the TED data and the evolutionary data from our functional families for drug repurposing and to predict novel enzymes in metagenomes.

10. Using AlphaFold2 for studying protein family evolution

Jessica Siltberg-Liberles, Hiram Duarte, and Kyoko Nakamura

Florida International University, Miami, Florida, USA

The AlphaFold Protein Structure Database (AFDB) provides extraordinary opportunities for studying protein evolution. We hypothesize that functionally diverse clades in the same protein family would also show divergence in structural properties. Here, the main findings for the conformationally flexible Calmodulin superfamily with several functionally divergent paralogs that mediate cellular signaling in response to changes in Ca2+ concentrations through allosteric effects and protein-protein interactions are presented. Maximum likelihood phylogenetic reconstruction of 448 proteins yielded 18 main clades. Functionally similar proteins from the same main clade reveal similar AlphaFold2 models and high model confidence. However, model variation is higher and model confidence lower for highly expanded plant-specific clades that lack functional annotations, display rapid sequence divergence and perhaps unexpected co-evolutionary patterns. Distance-based clustering based on structural similarity maintains the clades for most proteins even if the overall tree topology changed. The plant clades with several recent paralogs are dispersed across the tree, potentially caused by relaxed selective pressures following gene duplication events. AlphaFold2 models may inform our understanding of protein evolution including identifying functional divergence, especially if major conformations are represented by the models.

Comparisons of models of the Calmodulin superfamily proteins to their experimental structures found that proteins with high conformational flexibility show large differences, so some caution is warranted. Yet, our results support that AlphaFold2 can differentiate between functionally different clades and may detect patterns in individual sequences that cause different structural conformations and indirectly, functional divergence. To further establish strengths and weaknesses in using AlphaFold2 models for studying protein evolution on a large-scale level, we also investigated AlphaFold2 models from animals and plant proteins with different evolutionary history. The main findings from this work will be presented.

11. Addressing protein function and similarity via transformers.

Ana M. Rojas

Computational Biology and Bioinformatics, Andalusian Center for Developmental Biology (CABD-CSIC), Sevilla, Spain

We are entering the petabase era of sequence availability, where most sequences fall within the “twilight zone” of similarity. In such cases, traditional sequence annotation methods (largely dependent on the ortholog conjecture and evolutionary conservation) face increasing challenges. Functional protein annotation thus remains a major bottleneck for understanding the biology of both model and non-model organisms.

Recently, transformer-based models have emerged in computational biology, offering unprecedented potential for prediction tasks. In our work, we show that the ProtTrans model operates as a truly zero-shot predictor, successfully recovering functional signals from transcriptomic datasets when compared to traditional methods. Using this approach, we identified genes in 1,000 early non-model metazoans with functions consistent with each organism’s biology. To enable large-scale application, we developed a computational pipeline that integrates additional models.

Here, we explore how protein language model (pLM) embeddings relate to primary-sequence conservation by comparing embedding distances with aligned-sequence distances in a statistical framework. Applying this analysis to the RAS superfamily of proteins, we found that full-length protein embeddings accurately recapitulate classification in an alignment-free setting. These results suggest that pLM embeddings capture orthogonal functional features beyond simple residue conservation.

Altogether, our findings highlight the power of pLM-based annotation to expand functional insights in biodiversity research, while emphasizing the importance of interpreting embedding distances in the context of each model’s unique representational biases.

12. Covariation and coevolution: why the difference matters (to me)

David Talavera

The University of Manchester, Manchester, UK

The identification of covarying positions has been a major contributor to the development of AlphaFold, and it may lead to equivalent breakthroughs in the prediction of the interactome. Although it was usually assumed that observed patterns of covariation were mostly caused by

molecular coevolution, some findings provide a more nuanced picture. Previously, we found that many covarying pairs have low evolutionary rates and tend to be in buried in the core of the protein. Moreover, we showed that covarying substitutions mostly occur on different branches of the phylogenetic tree, casting doubt that coevolution is the main force for the observed covariation. These observations led us to propose the "coevolution paradox": The strength of coevolution required to cause coordinated changes means the evolutionary rate is so low that such changes are highly unlikely to occur.

More recently, we developed an approach for identifying coevolving positions based on maximum parsimony-based ancestral reconstruction followed by regression analyses. Our analyses show that the identified coevolving pairs tend to be close in the protein sequence and structure, slightly less solvent exposed and have a higher mutation rate than an equivalent random sample. Although some coevolving pairs have strong covariation, we demonstrate how this approach is essential for identifying pairs of coevolving positions with weak covariation patterns. Moreover, we do believe that ancestral reconstruction can also be used to detect favourable and unfavourable amino acid combinations, opening the door to the development of other Bioinformatics applications.

13. Covariant residue pairs reveals are important for enzyme activity but not stability

Matt Sternke, Katie Tripp, Soumya Behera, Justin Nguyen, & **Doug Barrick**

Department of Biophysics, Johns Hopkins University, Baltimore, USA

Pairwise correlations in multiple sequence alignments have long been known to identify structural contacts in proteins1,2 and is an important ingredient in AlphaFold prediction. One convenient way to extract pairwise biases from multiple sequence alignments is the Potts model, in which the energy of a protein sequence is represented by the sum of intrinsic single-site energies (represented with the letter h) and pairwise coupling energies (represented with the letter j). Previous work has shown that biochemical data from deep mutational scans (including stabilities, binding energies, kinetic parameters for enzyme catalysis, and cell growth) are better explained using models with pairwise coupling than models that include only single-site bias3,4.

To test the relative importance of pairwise (j) and single-site (h) bias on protein stability and activity, we have generated Potts coefficients from large MSAs from several protein families. We have used these coefficients to designed sequences with optimized Potts energies (□h + □j), sequences with limited pairwise terms, and sequences without pairwise terms (optimized only with single-site (□h) terms). Surprisingly, we find that sequences with optimized pairwise coupling (□h + □j), while more stable than extant proteins from the same families, are less stable than consensus proteins from the same alignments. Perhaps more surprisingly, sequences with optimized single-site (□h) terms are significantly more stable than consensus sequences. This observation is consistent over four unrelated protein families. For homeodomain, where we have characterized 62 different variants, correlation analysis reveals that stabilities are determined by single-site h terms, and that pairwise j terms have little to no effect on stability. These findings suggest that protein stability can be optimized by fitting a Potts model to an MSA and throwing out the j terms.

In contrast, measurements of enzyme kinetics for three different designed enzymes indicate that j optimization is essential for high turnover. Whether this is simply a result of a stability-activity tradeoff or of a more fundamental contribution of covariant residues to protein function remains to be seen.

(1) Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., & Weigt, M. Proc. Natl. Acad. Sci. U. S. A. 108, E1293-1301 (2011).
(2) Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Phys. Rev. E 87, 012707 (2013).
(3) Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Mol. Biol. Evol. 33, 268–280 (2016).
(4) Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Scharfe, C.P.I., Springer, M., Sander, C., Marks, D.S. Nat. Biotechnol. 35, 128–135 (2017).

14. Protein Evolution in Sequence Landscapes - From Data to Models and Back

Martin Weigt

Dept. of Computational, Quantitative and Synthetic Biology - CQSB, CNRS. Sorbonne Université, Paris, France

In the course of evolution, proteins diversify their sequences via a complex interplay between random mutations and neutral selection. As a consequence, we can today observe protein sequences of common evolutionary origin, with almost identical three-dimensional folds and biological functions, which however differ by as much as 70-80% of their amino acids. In my presentation, I will review our efforts to model protein evolution across multiple timescales, from the emergence of single mutations in a protein up to deep evolutionary time scales. To this aim, we first model protein fitness landscapes via generative probabilistic models trained on genomic data, and we show that these models are able to predict the effect of individual mutations, and to generate non-natural but biologically functional proteins. Second, we describe evolution as a stochastic process in these landscapes. The proposed framework accurately reproduces the sequence statistics of both short-time (experimental) and long-time (natural) protein evolution, suggesting applicability also to relatively data-poor intermediate evolutionary time scales, which are currently inaccessible to evolution experiments. Our model uncovers a highly collective nature of epistasis, gradually changing the fitness effect of mutations in a diverging sequence context, rather than acting via strong interactions between individual mutations. This collective nature triggers the emergence of a long evolutionary time scale, separating fast mutational processes inside a given sequence context, from the slow evolution of the context itself.

15. Coevolutionary contacts guiding the determination of alternative folds in proteins

Modesto Orozco

Institute for Research in Biomedicine (IRB), Barcelona, Spain

Co-evolutionary contacts, as detected through multiple sequence alignments, have become a cornerstone for protein structure prediction algorithms such as AlphaFold. Long before the advent of AlphaFold, we demonstrated that these contacts exhibit frustration—a signal that they may encode not just a single native fold, but information about alternative structural states. In this contribution, I will present our ongoing efforts to harness co-evolutionary signals

in combination with AlphaFold predictions and discrete molecular dynamics (DMD). Our goal is to uncover a hidden, dynamic landscape of the human proteome—one that reveals the potential for conformational variability and functional plasticity beyond static structural models.

16. Generative models learned on sequence data to forecast SARSCov2 viral evolution and antibody resilience

Simona Cocco

Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR8023 and Paris Sciences & Lettres (PSL) Research, Sorbonne Université, Paris, France

Generative models learned on sequence data to forecast SARS-CoV-2 viral evolution and antibody resilience. In this talk I will introduce the Restricted Boltzmann Machines (RBM): a simple Machine learning model with a bipartite graph architecture, learned only on sequence data. I will focus on the applications of RBM on the predictions of SARS-CoV-2 evolution. By integrating pre-pandemic evolutionary constraints gathered from SARS-CoV-2 far homologous sequences, with large-scale Deep mutational Scans (DMS) data we model how viral fitness, ACE2 binding, and immune escape pressures jointly sculpt the mutational landscape. Our sequence-based energy framework enables broad exploration of evolutionary trajectories while remaining valuable for experimental validation. Experimental validation of model predictions includes the test of 22 synthetic RBD variants with up to 21 mutations from the wild-type. Half of these variants maintained expression and ACE2 binding, and some successfully escaped most of the 9 antibodies tested.

17. The energetic structures of proteins

Ben Lehner

Centre for Genomic Regulation, Barcelona Spain

There are more ways to synthesise a 100-amino acid (aa) protein (20^100) than there are atoms in the universe. Only a very small fraction of such a vast sequence space can ever be experimentally or computationally surveyed. This lack of systematic experimental data limits our understanding of protein evolution. To address this shortcoming, we have experimentally sampled from sequence spaces larger than 10^10. Vast numbers of amino acid combinations constitute stable protein cores and surfaces. However, alternative cores frequently disrupt protein function by indirect allosteric effects. Fitting energy models to the data we have found that the genetic architecture of at least some proteins is strikingly simple, allowing accurate genetic prediction in high-dimensional sequence spaces with fully interpretable energy models. These models capture the nonlinear relationships between free energies and phenotypes but otherwise consist of additive free energy changes with a small contribution from pairwise energetic couplings. These energetic couplings are sparse and associated with structural contacts. Our results indicate that protein genetics is actually both rather simple and intelligible and suggest that allostery is an important constraint on sequence evolution.