Predicting protein 3D structure from evolutionary sequence variation

Thomas A. Hopf
Protein Prediction I, 06/18/2015
Outline

Prologue: Correlated mutations

Local vs. global models for 2D prediction

Application to 3D structure prediction
The structure prediction problem

genotype → phenotype

ACTGTGCACG
TAATGGGCATC
Structure from sequence alone?

Christian Anfinsen, Nobel Prize for Chemistry 1972
Sequence-structure gap is not closing!

Marks, Hopf & Sander, Nature Biotechnology (2012)
A protein
Evolutionary selection leaves residue covariation signature
Folding proteins from evolutionary couplings
Outline

Prologue: Correlated mutations

Local vs. global models for 2D prediction

Application to 3D structure prediction
Try something simple: correlation between two columns

single column frequencies: \( f_i(A_i) \) \( f_j(A_j) \)

column pair frequencies: \( f_{ij}(A_i, A_j) \)

\[ f_{ij}(A_i, A_j) - f_i(A_i) f_j(A_j) \]

To what extent do we see a pair of amino acids more/less often than expected by chance?
Mutual information measures correlation between two columns. If the empirical frequency distribution of amino acids A and B between two columns i and j is independent, we would expect the product of the individual frequency distributions $f_{imA} f_{jnB}$ to be roughly equal to the joint frequency distribution $f_{ijmAB}$, while a deviation indicates a correlation between the two columns.

### 2.3.2 Sequence weighting

Frequency counts as defined in Equations mwsvn and mwswn are prone to uneven sampling of sequence space due to experimental bias. To decrease the influence of sampling bias on the estimation of correlations, sequences in the multiple alignment are downweighted based on the number of neighbors with sequence identity above a similarity threshold $\mu < \mu'$. Formally, for each sequence $m$ we calculate the number of sequences in the alignment including $m$ $n$ which have at least $L$ aligned identical residues as:

$$k_m = \sum_{i=1}^{L} \sum_{n=1}^{M} m_{wsix}$$

Assigning a weight of $v/k_m$ to each sequence $m$, the frequency counts from Equations mwsvn and mwswn can be redefined as $f_{imA} = \frac{v}{k_m} f_{wsix}$ and $f_{ijmAB} = \frac{v}{k_m} f_{wsxn}$, respectively. In both cases, $M_e = \sum_{m=1}^{M} 1/k_m$ is the effective number of sequences in the alignment after weighting, and $\mu$ is a pseudocount term to avoid non-zero entries in the correlation matrix defined in Section wsxns.

Initial tests by Marks et al. [yy] and Morcos et al. [zx] suggest a choice of $\mu = u.7$ and $\mu' = u.5$ for reasonable performances.

### 2.3.3 Calculation of direct information

A commonly used measure for calculating the correlation between two columns in a sequence alignment is mutual information $MI$, which is defined as:

$$MI_{ij} = \sum_{A_i, A_j=1}^{q} f_{ij}(A_i, A_j) \ln \left( \frac{f_{ij}(A_i, A_j)}{f_i(A_i)f_j(A_j)} \right)$$

- **sum all possible amino acid combinations**
- **weight**
- **deviation from statistical independence**
Bad news: doesn’t work.

local model
main problem: transitivity

PDB structure residue contacts
residue pairs with 100 highest MI values
Solution: use a global model!

Global probability model explains observed correlations by causative pair interactions.

Inverse Potts inference (undirected graphical model)

Observed correlations: causal, direct coevolution.
Probability model connects correlations to direct interactions

\[ P(\sigma) = \frac{1}{Z} \exp \left( \sum_{i=1}^{N} h_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(\sigma_i, \sigma_j) \right) \]

observed data (sequences) → approximate maximum likelihood inference → direct pair interactions
From sequences to pair scores

Infer parameters

\[ P(\text{data} | \text{parameters}) = \prod_{\sigma \in \text{alignment}} P(\sigma | h, J) \]

approximate!

Calculate evolutionary couplings

\[ FN_{ij} = \| J_{ij} \|_2 = \sqrt{\sum_{k=1}^{21} \sum_{k=1}^{21} J_{ij}(k, l)^2} \]

+ some other technical details

References: Marks et al. (2011), Ekeberg et al. (2013)
Most global model pairs are close in 3D

local model
(mutual information)

global model
(MaxEnt, Marks et al., 2011)
He solved it before everyone else did...

Alan Lapedes
Outline

Prologue: Correlated mutations

Local vs. global models for 2D prediction

Application to 3D structure prediction
What could we predict using evolutionary couplings?

3D structure

function

protein complexes

conformation changes
What could we predict?

3D structure

protein complexes
A brief reminder
The breakthrough: 15 proteins folded from sequences alone

Marks et al., PLoS ONE (2011)
Also works for membrane proteins!

predicted  experimental

Hopf et al., Cell (2012)
Medically important membrane proteins predicted from sequences

predicted
(Hopf et al., Cell, 2012)

experimental
(Baradaran et al., Nature, 2013)

respiratory complex I
Medically important membrane proteins predicted from sequences II

human adiponectin receptor 1
(3.7Å, 186 residues)
What could we predict?

3D structure

protein complexes
Interacting proteins co-evolve to maintain interaction
Complex interactions from the evolutionary sequence record
Accurate prediction of the ABC transporter MetNI
Benchmark set gallery

CyoA – CyoB
EnvZ – OmpR (homolog)
MoaD – MoaE
FimC – FimD
BtuC – BtuF
BtuC – BtuD
DhaK – DhaL
CarB – CarA
GcsT – GcsH
RS3 – RS14
De novo prediction of unsolved complex interactions
Can we predict which proteins interact?

E. coli ATP synthase
ATP synthase interaction matrix
Try folding yourself!

www.evfold.org
References for further reading

Protein 3D Structure Computed from Evolutionary Sequence Variation

Debora S. Marks¹, Lucy J. Colwell³, Robert Sheridan⁷, Thomas A. Hopf¹, Andrea Pagnani⁵, Riccardo Zecchina⁴, Chris Sander²

¹Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, United States of America, ²MRG Laboratory of Molecular Biology, Hills Road, Cambridge, United Kingdom, ³Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America, ⁴Human Genetics Foundation, Torino, Italy, ⁵Politecnico di Torino, Torino, Italy

Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing

Thomas A. Hopf¹,² Lucy J. Colwell³, Robert Sheridan⁷, Burkhard Rost⁴, Chris Sander⁴ and Debora S. Marks¹,⁴

¹Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA, ²Department of Informatics, Technische Universität München, IS148 Garching, Germany, ³MRG Laboratory of Molecular Biology, Hills Road, CB2 0XH Cambridge, UK, ⁴Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York City, NY 10065, USA

*Correspondence: av48@memorial Sloan-kettering.org

DOI 10.1371/journal.pone.0028766

Protein structure prediction from sequence variation

Debora S Marks¹, Thomas A Hopf⁴ & Chris Sander²

Sequence co-evolution gives 3D contacts and structures of protein complexes

Thomas A Hopf¹,², Charlotte P I Schärfe³,⁴, Joao P G L M Rodrigues⁵, Anna G Green¹, Oliver Kolbacher²,³, Chris Sander⁴, Alexandre M J J Bonvin⁵, Debora S Marks¹
Acknowledgments

Debora Marks  
(Harvard Medical School)

Chris Sander  
(MSKCC)

Burkhard Rost  
(TU München)

Charlotta Schärfe  
(HMS & Uni Tübingen)