THE FRANCIS CRICK INSTITUTE

Using Deep Learning to Predict Protein Structure and Function

Prof. David Jones
UCL & The Francis Crick Institute

MRC Medical Research Council

UCL

Imperial College London

KING'S College LONDON

wellcome

CANCER RESEARCH UK

THE FRANCIS CRICK INSTITUTE

The Francis Crick Institute

Deep Learning

Recent Improvements

- 2006, Deep Belief Networks
- 2011, Rectified Linear Units
- 2014, Dropout technique
- 2015, Batch Normalization

- 2010, THEANO (Acceleration by GPUs, development of software frameworks)
- Cloud access to GPU/TPU hardware

Open Source Resources

theano

Lasagne

PYTORCH

Caffe

cuDNN

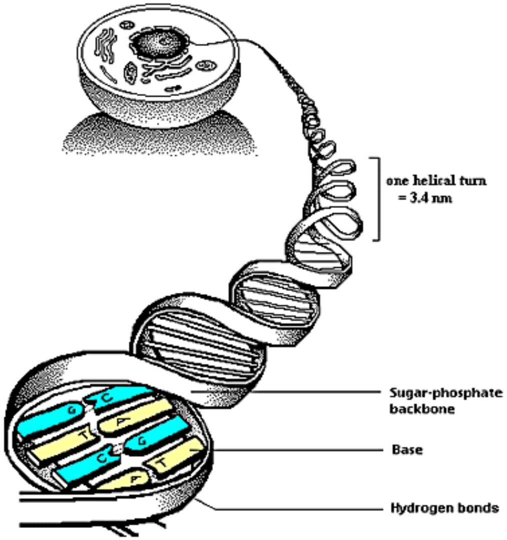
TensorFlow

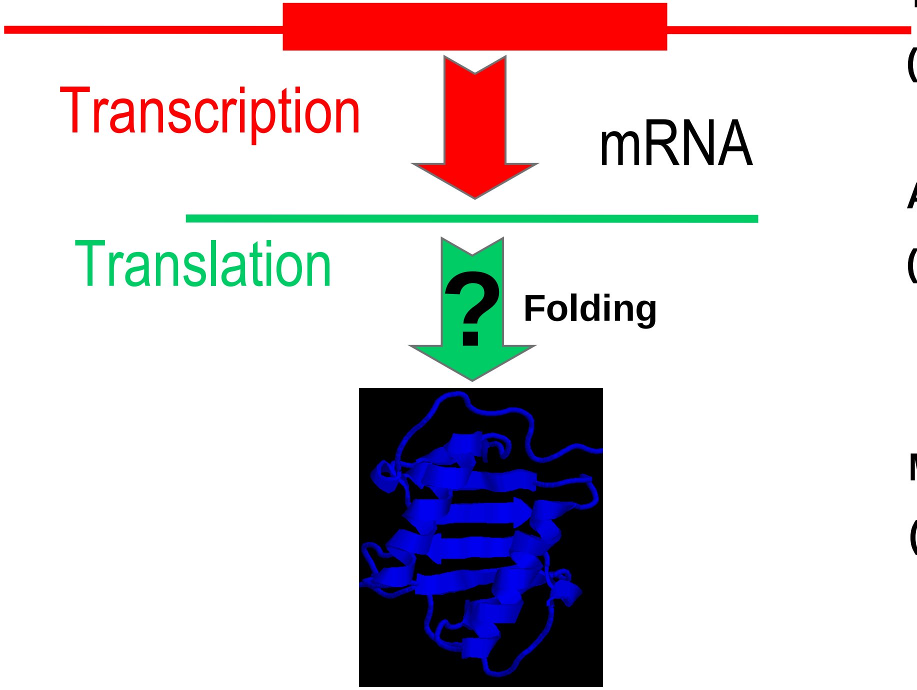
K

Deep Learning Tutorials:
deeplearning.net/tutorial/
deeplearning4j.org

torch

The Central Dogma of Molecular Biology





DNA

Gene

Transcription

mRNA

Translation

?

Folding

Folded Functional Protein

TATGCTGAGGTA…

(4 letter code)

AUACGACUCCAU…

(4 letter code)

MLLPVTCAYVKGKLQ…

(20 letter code)

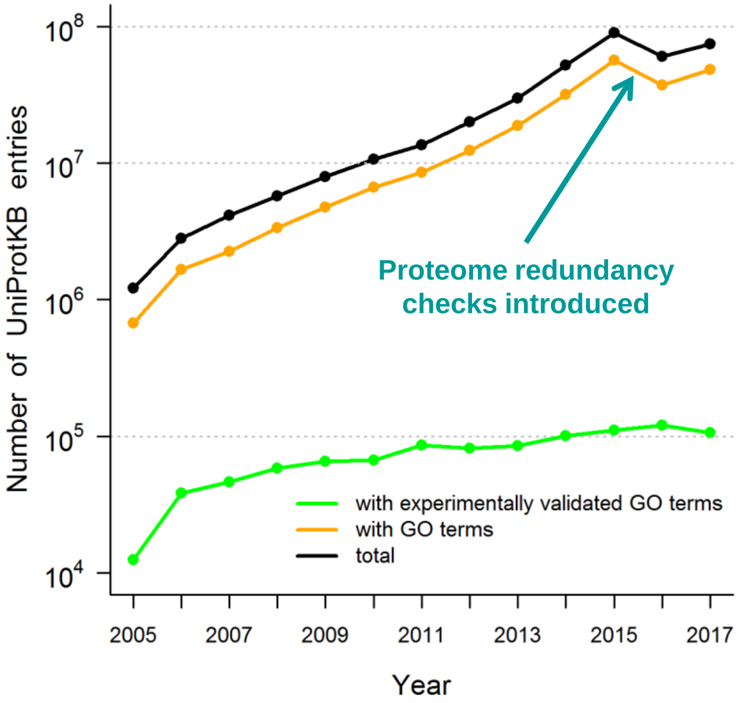
Current state of protein function annotation

- The function of the vast majority of sequences in UniProtKB has not been experimentally annotated yet

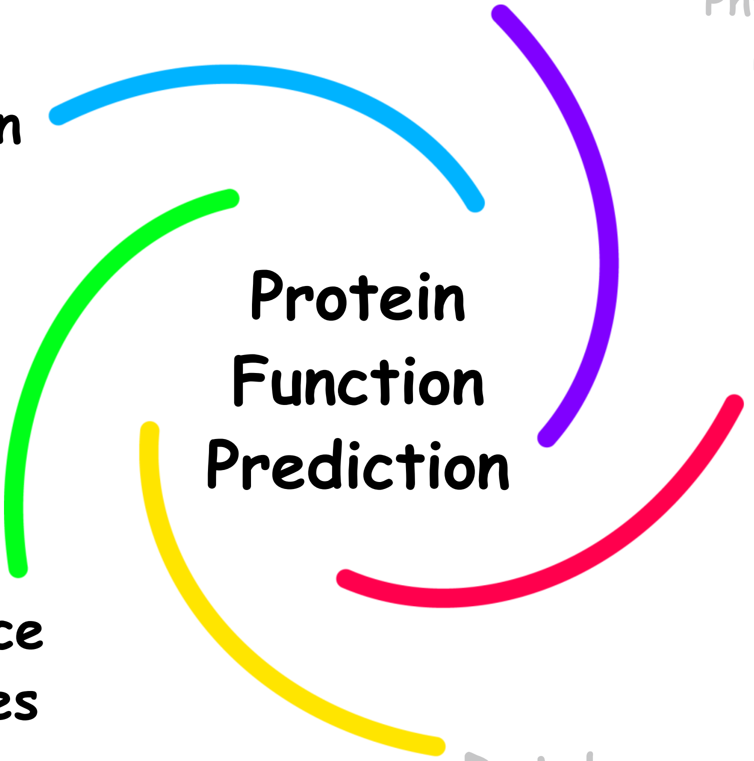
- Standard methods for annotation transfers from sequence allow to fill in this gap only partially

- About 40% of human chains in UniProtKB still have no GO term assignments

- And about 60% lack information for the biological process (BP) domain



THE FRANCIS CRICK INSTITUTE



Protein Function Prediction

Protein-protein Interactions

Domain Fusion

Phylogenetic Profiles

Gene Expression

Evolutionary Information

Domain Architecture

Homologous Relationships

Orthologous Relationships

Database or literature annotations

Sequence Features

Secondary Structure

Disordered Regions

Transmembrane Segments

Signal Peptides

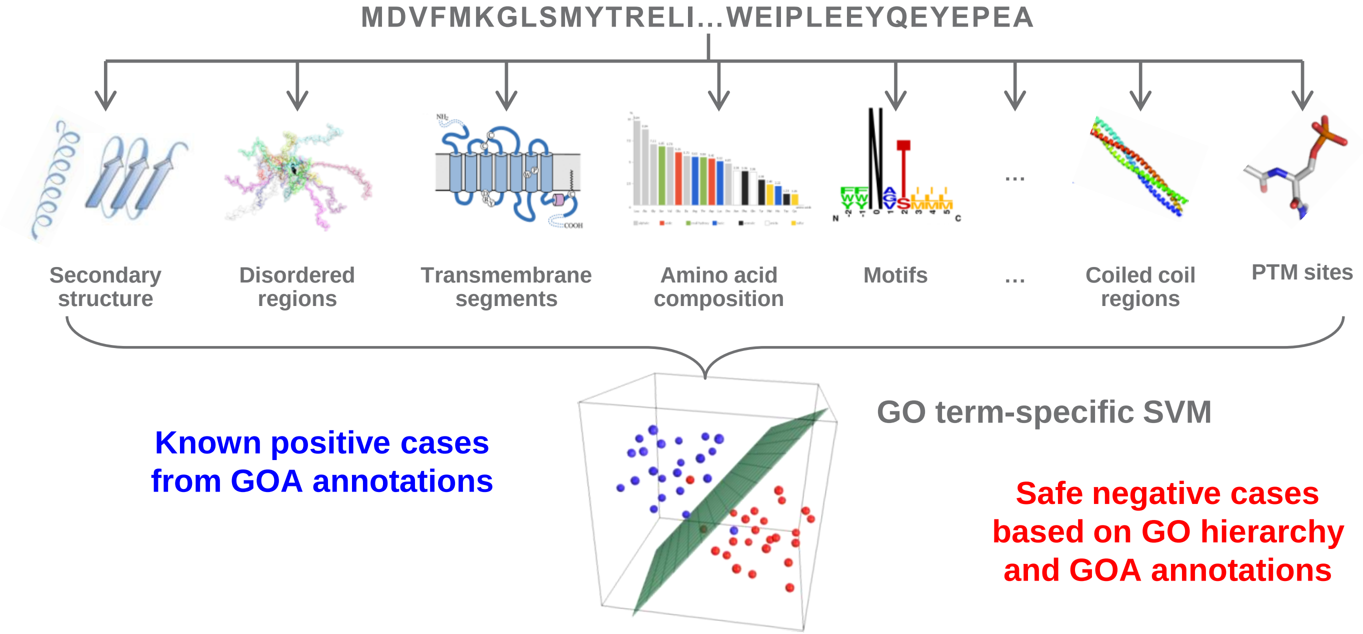
AA Composition

PTM Sites

De novo function prediction from sequence

Approach pioneered in Soren Brunak’s lab with **ProtFun**

FFPred adds protein intrinsic disorder information to the array of features



MDVFMKGLSMYTRELI…WEIPLEEYQEYEPEA

Secondary structure

Disordered regions

Transmembrane segments

Amino acid composition

Motifs

…

Coiled coil regions

PTM sites

GO term-specific SVM

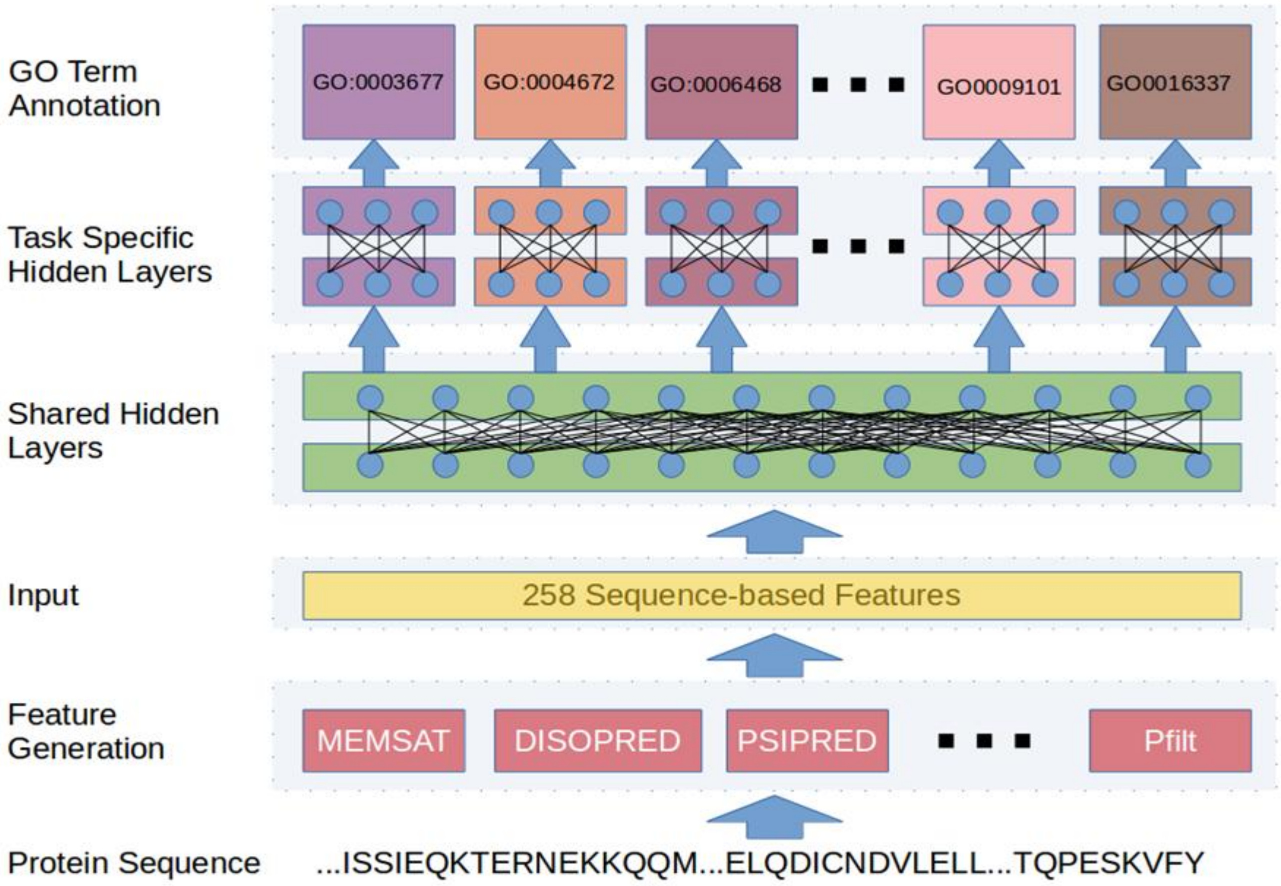
Known positive cases from GOA annotations

Safe negative cases based on GO hierarchy and GOA annotations

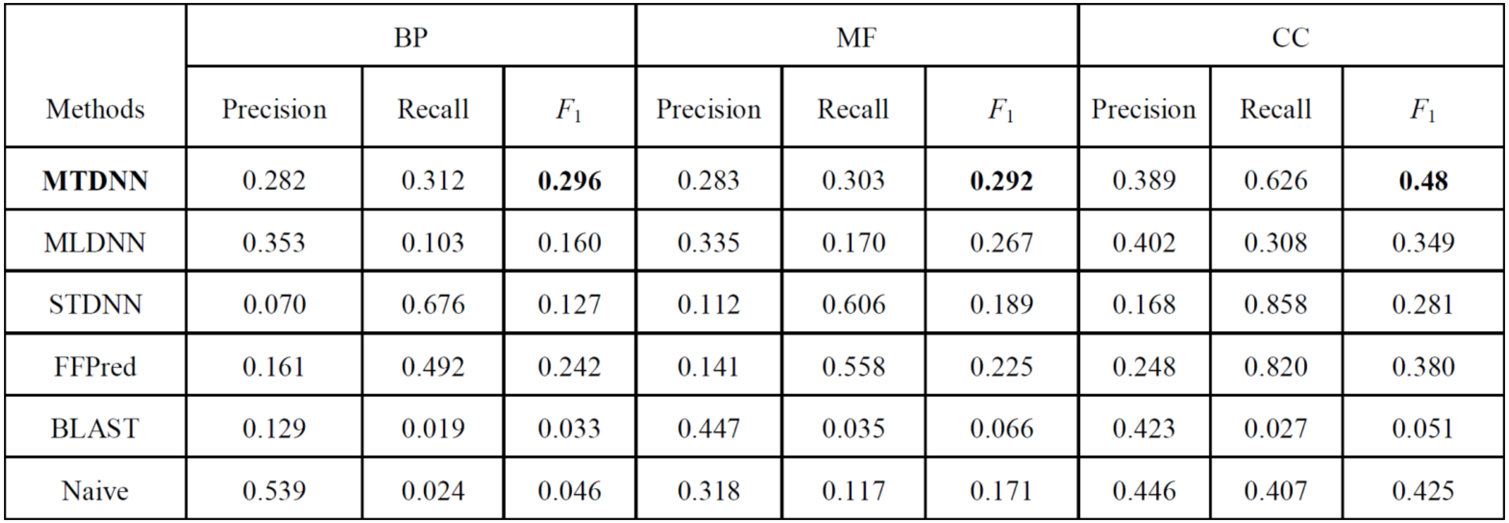
Multi-Class, Multi-Label and Multi-Task Learning

- Multi-Class: instances are classified in just one of more than two classes, like secondary structure prediction;
- Multi-Label: multiple target labels may be assigned to each instance, like protein function prediction;
- In a structural point of view, Multi-Task methods learn different tasks in parallel while using a single shared representation.
- Using a Multi-Task structure to solve the Multi-Label problem is useful because different GO term labels may have wildly different sized positive and negative sets for training.

Multi-Task Deep Neural Networks



Multitask DNNs Benchmark Results



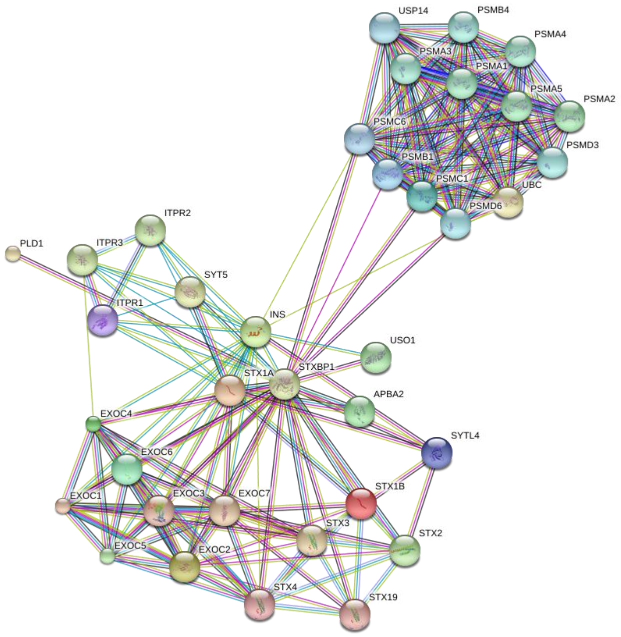
Methods	BP			MF			CC		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
MTDNN	0.282	0.312	**0.296**	0.283	0.303	**0.292**	0.389	0.626	**0.48**
MLDNN	0.353	0.103	0.160	0.335	0.170	0.267	0.402	0.308	0.349
STDNN	0.070	0.676	0.127	0.112	0.606	0.189	0.168	0.858	0.281
FFPred	0.161	0.492	0.242	0.141	0.558	0.225	0.248	0.820	0.380
BLAST	0.129	0.019	0.033	0.447	0.035	0.066	0.423	0.027	0.051
Naive	0.539	0.024	0.046	0.318	0.117	0.171	0.446	0.407	0.425

Learning Protein Function from Biological networks

- Biological networks underlie all aspects of protein function and are a convenient model to represent, analyse and reason about protein-protein interactions (functional links) from known and predicted data sources..

- The STRING database includes associations describing:
    - experimentally detected interactions;
    - conserved mRNA co-expression;
    - conserved gene proximity;
    - co-mention in abstracts and papers;
    - interactions from curated databases;
    - gene co-occurrence/co-absence;
    - gene fusion events.

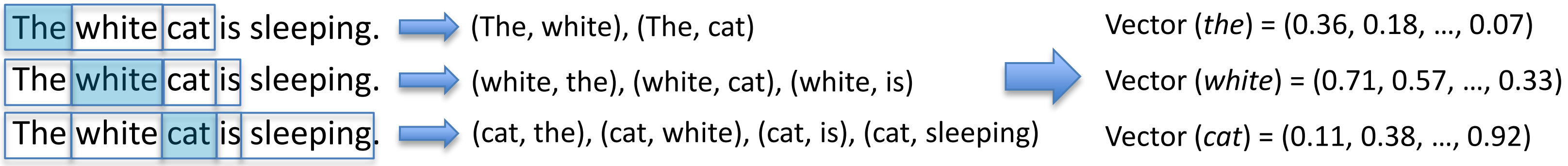
- These heterogeneous data are combined through naïve Bayes statistics



From word embedding to graph embedding

Word Embedding - Word2Vec (the Skip-Gram model)

- IDEA: Learning context feature for the target word by maximising the probability of neighboring word co-occurrence.
- A shallow neural network architecture feature learning method.



The white cat is sleeping. → (The, white), (The, cat)

The white cat is sleeping. → (white, the), (white, cat), (white, is)

The white cat is sleeping. → (cat, the), (cat, white), (cat, is), (cat, sleeping)

→

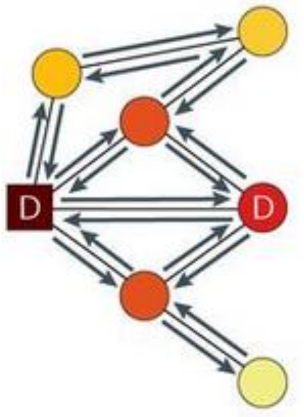
Vector (*the*) = (0.36, 0.18, …, 0.07)

Vector (*white*) = (0.71, 0.57, …, 0.33)

Vector (*cat*) = (0.11, 0.38, …, 0.92)

Graph Embedding - Node2Vec

- Extension from Word2Vec
- Using Random-walks to learn the context of network
  (a node in the network = a word in a sentence)
- Can be applied to arbitrary graph/network



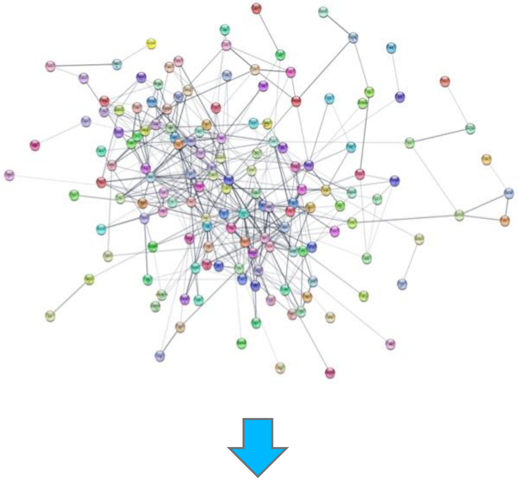
Random-walk on a graph

Mashup/Node2Vec – learning representations of network topology

- It explores the network through **random walks** to learn the diffusion states of network or generate sequences of nodes.

- It learns a feature space that optimally approximates the original node diffusion states or preserves maximum node neighborhood information.

- Similar network embedding algorithms have been proposed: Deep-walk [3], Line [4]

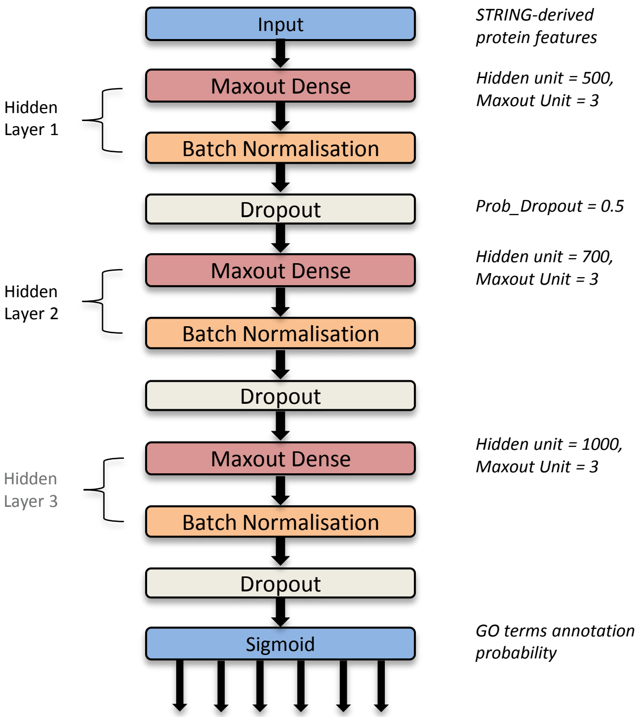


N-dimensional vectors for
each node (protein)

[1] Cho, H., Berger, B., & Peng, J. Compact integration of multi-network topology for functional analysis of
genes. Cell systems . 2016
[2] Grover, A., & Leskovec, J., node2vec: Scalable feature learning for networks. *KDD 2016*.
[3] Perozzi, B., Al-Rfou, R., & Skiena, S., Deepwalk: Online learning of social representations. *KDD 2014*.
[4] Tang, J. et. al, Line: Large-scale information network embedding. *WWW 2015*.

Maxout Deep Neural Networks

- A simple fully-connected three-hidden-layer neural network
- Input Node2Vec generated features into Deep Maxout Neural Networks (DMNN)
- Output multiple GO term annotations for each protein
- Hidden neurons: 300, 500, 700 or 1000
- Maxout units: 3 per layer
- Batch normalisation
- Dropout - Prob. = 0.5

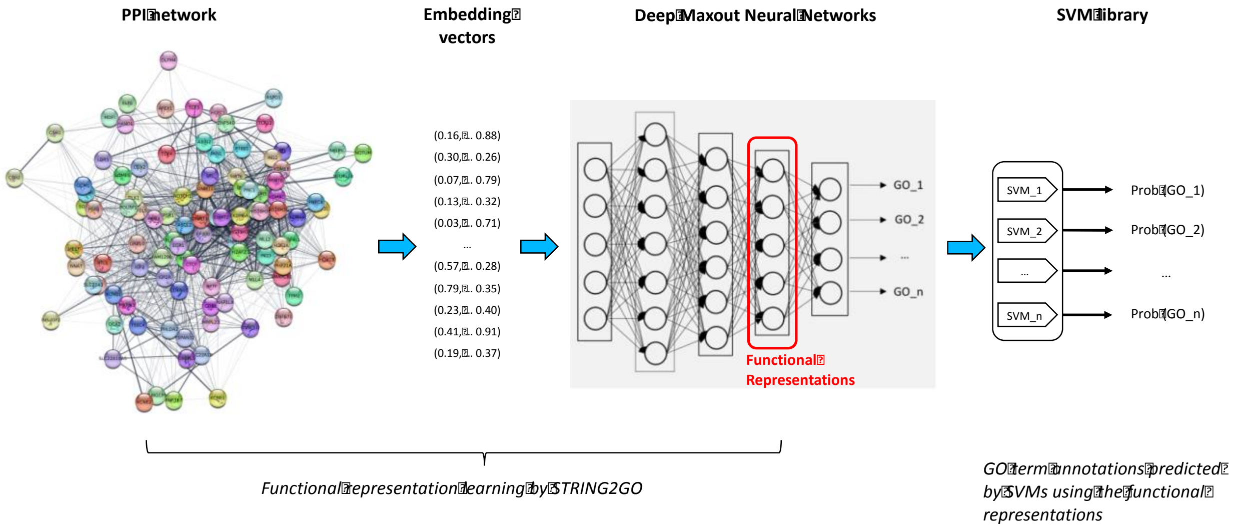


Network architecture

[1] Goodfellow, I. J., et al. Maxout networks. *arXiv 2013.*
[2] Ioffe, S., & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML 2015*.
[3] Srivastava, N., et al. Dropout: a simple way to prevent neural networks from overfitting. *JMLR 2014.*

STRING2GO: Predicting GO terms using DNN functional representations and a Support Vector Machine Output Layer

- Use Deep Maxout Neural Networks to learn the functional representations but NOT to actually assign the labels
- Train the network exactly as before but replace the multi-label output layer with a library of binary SVM classifiers



PPI network

Embedding vectors

(0.16, .. 0.88)
(0.30, .. 0.26)
(0.07, .. 0.79)
(0.13, .. 0.32)
(0.03, .. 0.71)
...
(0.57, .. 0.28)
(0.79, .. 0.35)
(0.23, .. 0.40)
(0.41, .. 0.91)
(0.19, .. 0.37)

Deep Maxout Neural Networks

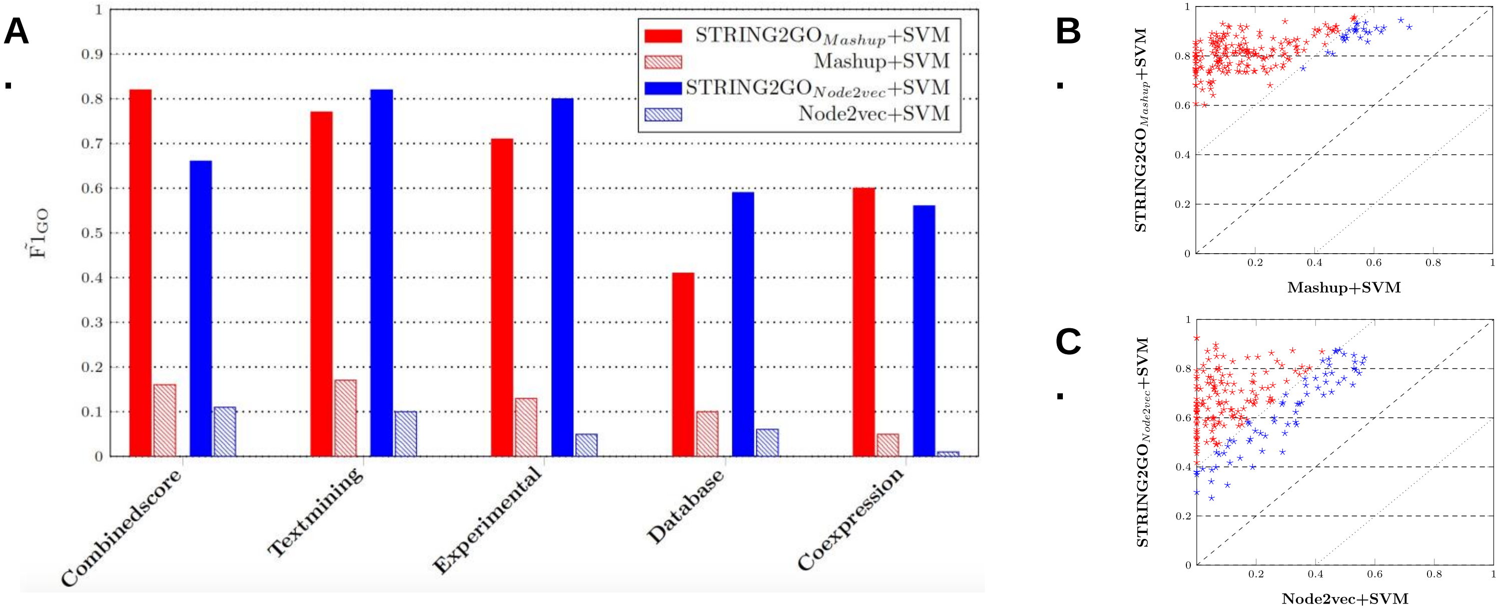
Functional Representations

SVM library

Functional representation learning by STRING2GO

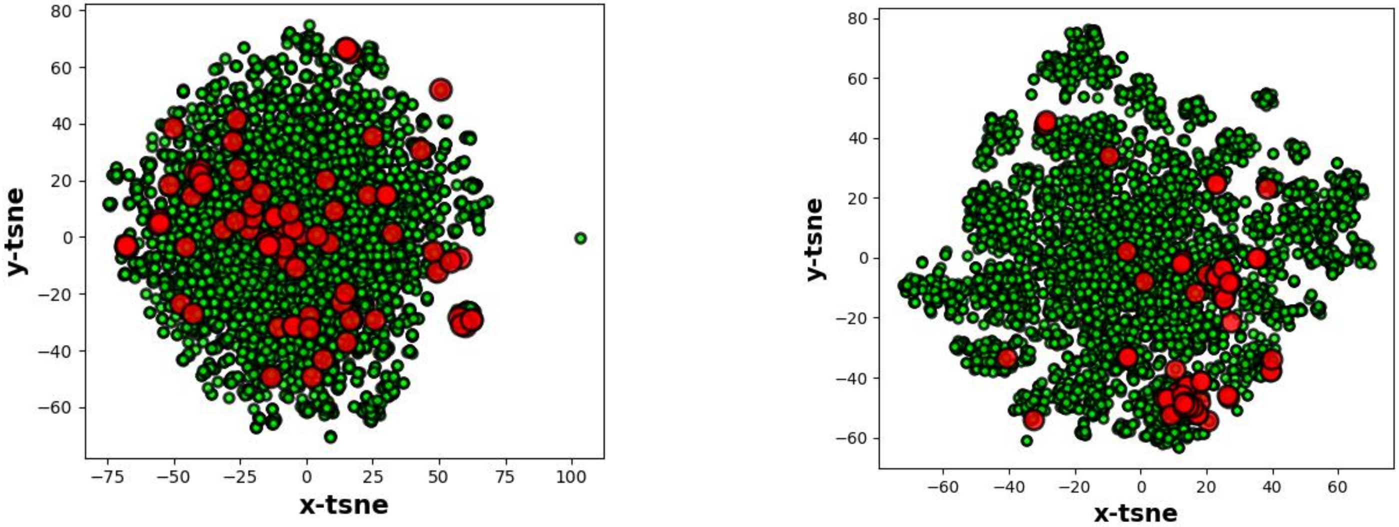
GO term annotations predicted by SVMs using the functional representations

Learned latent functional representations greatly outperform raw network-embedding features in function prediction



(A) Median F1 score for predicting 204 BP terms over 10-fold cross validation during training stage using different STRING networks to generate network-embedding features and corresponding functional representations (B,C) F1 score obtained by different features based on Combinedscore network

DNN latent functional representation offer enhanced functional discrimination power on protein test cases



(A) **(B)**

tsne = t-distributed stochastic neighbor embedding

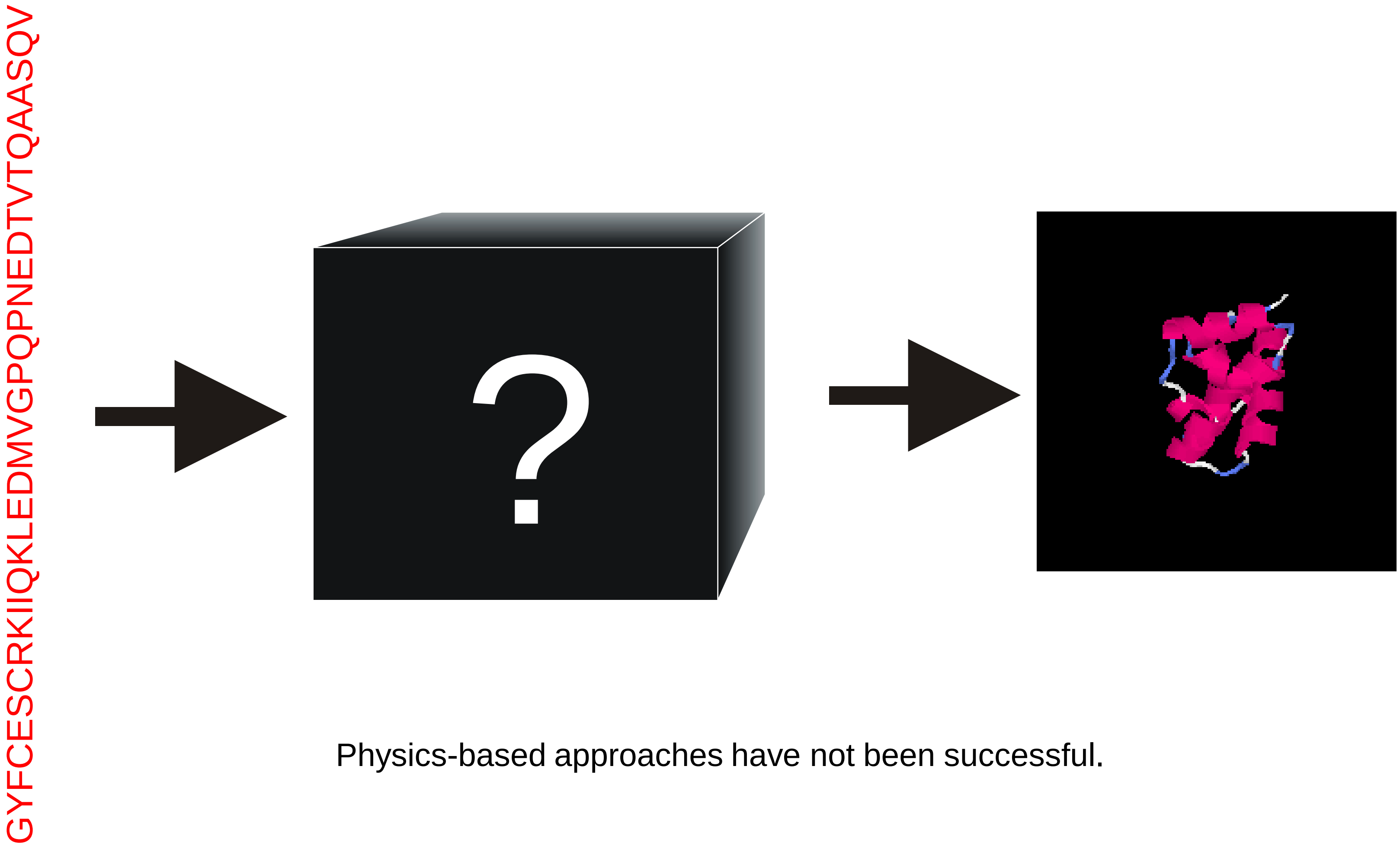
Colored protein samples labelled by GO:0090150 - annotated (red) or non-annotated (green) - using raw Mashup-derived features (A) and corresponding functional representations (B)

The Protein Folding Problem

A machine learning approach

THE FRANCIS CRICK INSTITUTE

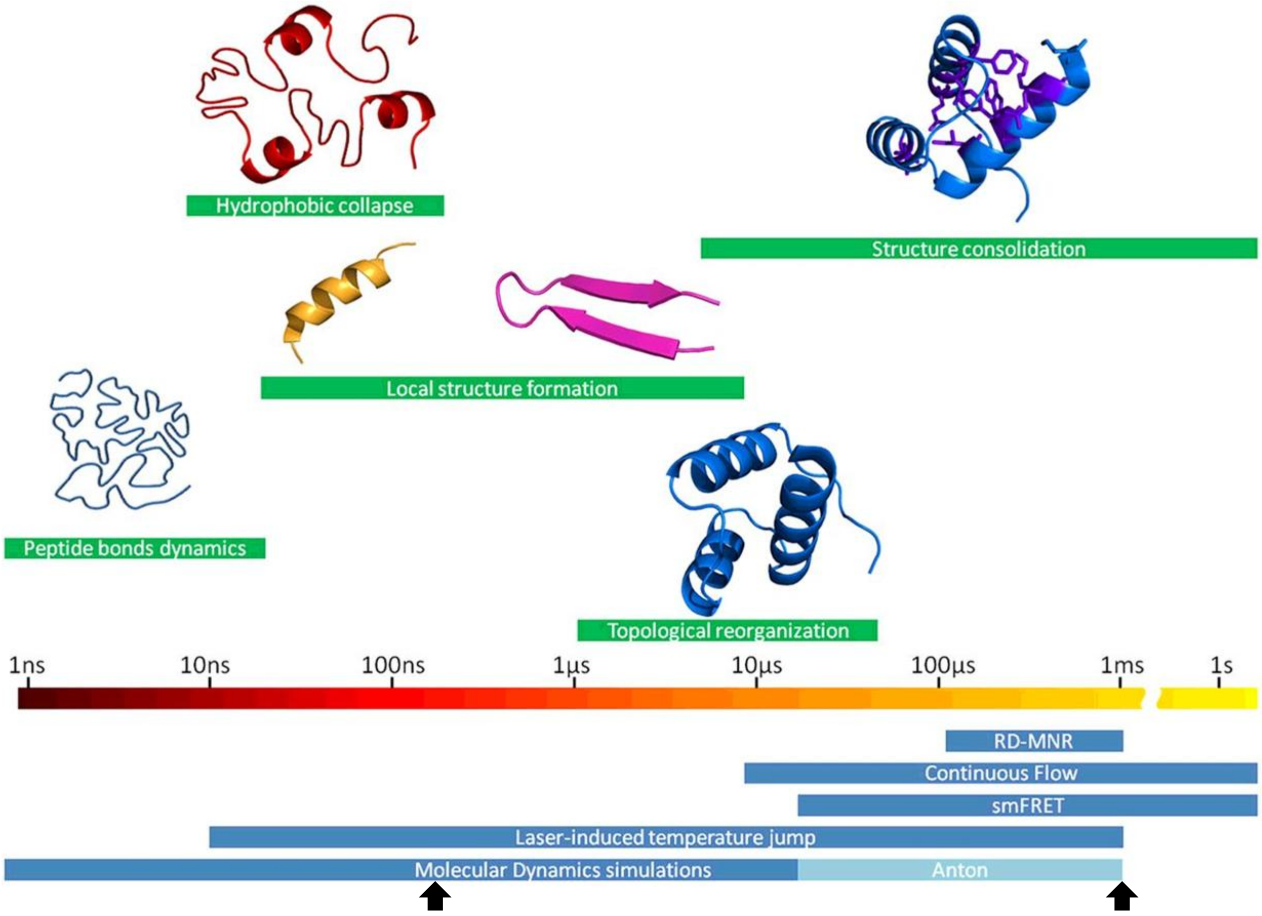
A Grand Challenge in Biology for 50 Years: The Protein Folding Problem



GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQV

Physics-based approaches have not been successful.

Protein Folding Timescales

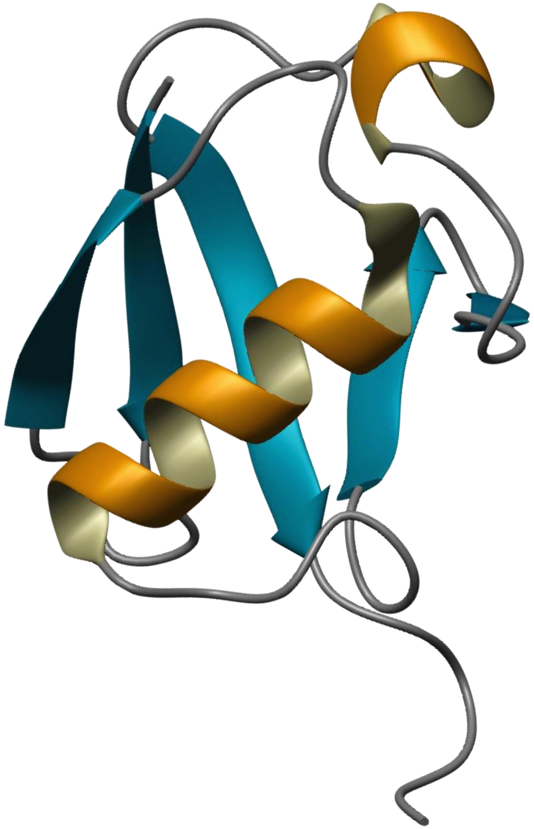


1 CPU day

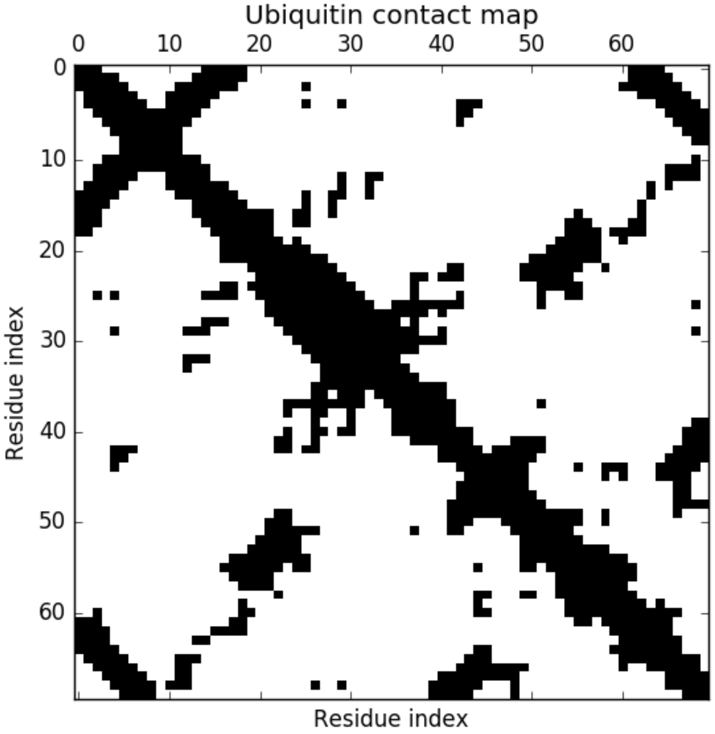
16 CPU years

Victor Muñoz, and Michele Cerminara Biochem. J. 2016;473:2545-2559

The Protein Folding Problem as a Graphical Modelling Problem

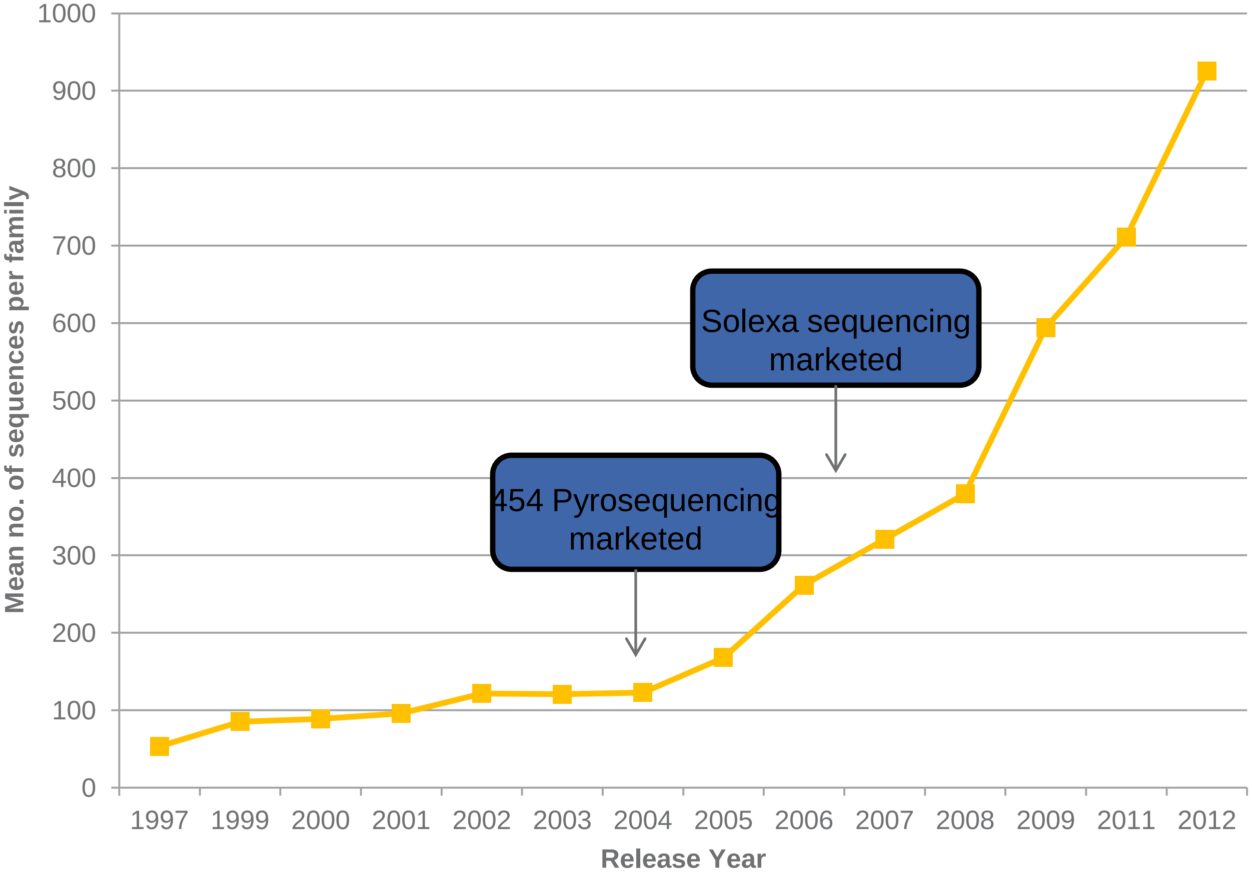


Calculate distances from coordinates and apply a threshold



The contact map can be thought of as an adjacency matrix.

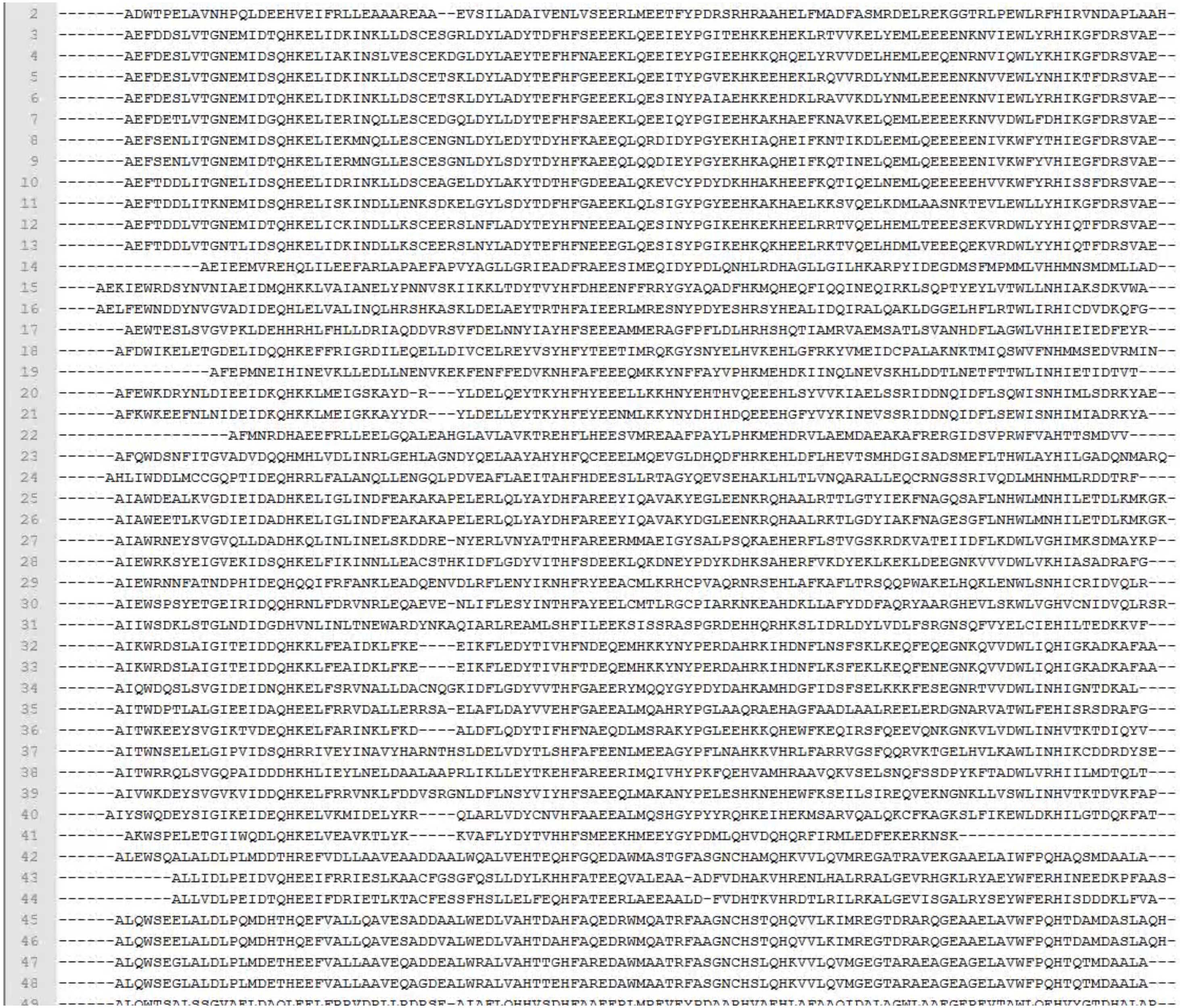
Growth of Sequence Family Size in Pfam



Solexa sequencing marketed

454 Pyrosequencing marketed

[redacted]



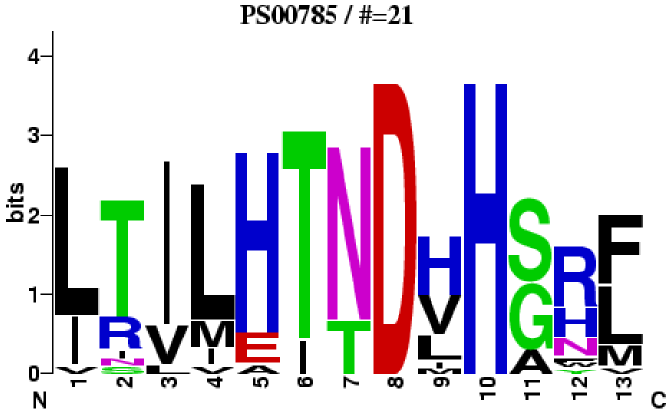
```
 2  -------ADWTPELAVNHPQLDEEHVEIFRLLEAAAREAA--EVSILADAIVENLVSEERLMEETFYPDRSRHRAAHELFMADFASMRDELREKGGTRLPEWLRFHIRVNDAPLAAH-
 3  -------AEFDDSLVTGNEMIDTQHKELIDKINKLLDSCESGRLDYLADYTDFHFSEEEKLQEEIEYPGITEHKKEHEKLRTVVKELYEMLEEEENKNVIEWLYRHIKGFDRSVAE--
 4  -------AEFDESLVTGNEMIDSQHKELIAKINSLVESCEKDGLDYLAEYTEFHFNAEEKLQEEIEYPGIEEHKKQHQELYRVVDELHEMLEEQENRNVIQWLYKHIKGFDRSVAE--
 5  -------AEFDESLVTGNEMIDSQHKELIDKINKLLDSCETSKLDYLADYTEFHFGEEEKLQEEITYPGVEKHKEEHEKLRQVVRDLYNMLEEEENKNVVEWLYNHIKTFDRSVAE--
 6  -------AEFDESLVTGNEMIDTQHKELIDKINKLLDSCETSKLDYLADYTEFHFGEEEKLQESINYPAIAEHKKEHDKLRAVVKDLYNMLEEEENKNVIEWLYRHIKGFDRSVAE--
 7  -------AEFDETLVTGNEMIDGQHKELIERINQLLESCEDGQLDYLLDYTEFHFSAEEKLQEEIQYPGIEEHKAKHAEFKNAVKELQEMLEEEEKKNVVDWLFDHIKGFDRSVAE--
 8  -------AEFSENLITGNEMIDSQHKELIEKMNQLLESCENGNLDYLEDYTDYHFKAEEQLQRDIDYPGYEKHIAQHEIFKNTIKDLEEMLQEEEEENIVKWFYTHIEGFDRSVAE--
 9  -------AEFSENLVTGNEMIDTQHKELIERMNGLLESCESGNLDYLSDYTDYHFKAEEQLQQDIEYPGYEKHKAQHEIFKQTINELQEMLQEEEEENIVKWFYVHIEGFDRSVAE--
10  -------AEFTDDLITGNELIDSQHEELIDRINKLLDSCEAGELDYLAKYTDTHFGDEEALQKEVCYPDYDKHHAKHEEFKQTIQELNEMLQEEEEEHVVKWFYRHISSFDRSVAE--
11  -------AEFTDDLITKNEMIDSQHRELISKINDLLENKSDKELGYLSDYTDFHFGAEEKLQLSIGYPGYEEHKAKHAELKKSVQELKDMLAASNKTEVLEWLLYHIKGFDRSVAE--
12  -------AEFTDDLVTGNEMIDTQHKELICKINDLLKSCEERSLNFLADYTEYHFNEEEALQESINYPGIKEHKEKHEELRRTVQELHEMLTEEESEKVRDWLYYHIQTFDRSVAE--
13  -------AEFTDDLVTGNTLIDSQHKELIDKINDLLKSCEERSLNYLADYTEFHFNEEEGLQESISYPGIKEHKQKHEELRKTVQELHDMLVEEEQEKVRDWLYYHIQTFDRSVAE--
14  ---------------AEIEEMVREHQLILEEFARLAPAEFAPVYAGLLGRIEADFRAEESIMEQIDYPDLQNHLRDHAGLLGILHKARPYIDEGDMSFMPMMLVHHMNSMDMLLAD--
15  ----AEKIEWRDSYNVNIAEIDMQHKKLVAIANELYPNNVSKIIKKLTDYTVYHFDHEENFFRRYGYAQADFHKMQHEQFIQQINEQIRKLSQPTYEYLVTWLLNHIAKSDKVWA---
16  ----AELFEWNDDYNVGVADIDEQHLELVALINQLHRSHKASKLDELAEYTRTHFAIEERLMRESNYPDYESHRSYHEALIDQIRALQAKLDGGELHFLRTWLIRHICDVDKQFG---
17  -------AEWTESLSVGVPKLDEHHRHLFHLLDRIAQDDVRSVFDELNNYIAYHFSEEEAMMERAGFPFLDLHRHSHQTIAMRVAEMSATLSVANHDFLAGWLVHHIEIEDFEYR---
18  ------AFDWIKELETGDELIDQQHKEFFRIGRDILEQELLDIVCELREYVSYHFYTEETIMRQKGYSNYELHVKEHLGFRKYVMEIDCPALAKNKTMIQSWVFNHMMSEDVRMIN--
19  ----------------AFEPMNEIHINEVKLLEDLLNENVKEKFENFFEDVKNHFAFEEEQMKKYNFFAYVPHKMEHDKIINQLNEVSKHLDDTLNETFTTWLINHIETIDTVT----
20  ------AFEWKDRYNLDIEEIDKQHKKLMEIGSKAYD-R---YLDELQEYTKYHFHYEEELLKKHNYEHTHVQEEEHLSYVVKIAELSSRIDDNQIDFLSQWISNHIMLSDRKYAE--
21  ------AFKWKEEFNLNIDEIDKQHKKLMEIGKKAYYDR---YLDELLEYTKYHFEYEENMLKKYNYDHIHDQEEEHGFYVYKINEVSSRIDDNQIDFLSEWISNHIMIADRKYA---
22  ------------------AFMNRDHAEEFRLLEELGQALEAHGLAVLAVKTREHFLHEESVMREAAFPAYLPHKMEHDRVLAEMDAEAKAFRERGIDSVPRWFVAHTTSMDVV-----
23  ------AFQWDSNFITGVADVDQQHMHLVDLINRLGEHLAGNDYQELAAYAHYHFQCEEELMQEVGLDHQDFHRKEHLDFLHEVTSMHDGISADSMEFLTHWLAYHILGADQNMARQ-
24  -----AHLIWDDLMCCGQPTIDEQHRRLFALANQLLENGQLPDVEAFLAEITAHFHDEESLLRTAGYQEVSEHAKLHLTLVNQARALLEQCRNGSSRIVQDLMHNHMLRDDTRF----
25  ------AIAWDEALKVGDIEIDADHKELIGLINDFEAKAKAPELERLQLYAYDHFAREEYIQAVAKYEGLEENKRQHAALRTTLGTYIEKFNAGQSAFLNHWLMNHILETDLKMKGK-
26  ------AIAWEETLKVGDIEIDADHKELIGLINDFEAKAKAPELERLQLYAYDHFAREEYIQAVAKYDGLEENKRQHAALRKTLGDYIAKFNAGESGFLNHWLMNHILETDLKMKGK-
27  ------AIAWRNEYSVGVQLLDADHKQLINLINELSKDDRE-NYERLVNYATTHFAREERMMAEIGYSALPSQKAEHERFLSTVGSKRDKVATEIIDFLKDWLVGHIMKSDMAYKP--
28  ------AIEWRKSYEIGVEKIDSQHKELFIKINNLLEACSTHKIDFLGDYVITHFSDEEKLQKDNEYPDYKDHKSAHERFVKDYEKLKEKLDEEGNKVVVDWLVKHIASADRAFG---
29  ------AIEWRNNFATNDPHIDEQHQQIFRFANKLEADQENVDLRFLENYIKNHFRYEEACMLKRHCPVAQRNRSEHLAFKAFLTRSQQPWAKELHQKLENWLSNHICRIDVQLR---
30  ------AIEWSPSYETGEIRIDQQHRNLFDRVNRLEQAEVE-NLIFLESYINTHFAYEELCMTLRGCPIARKNKEAHDKLLAFYDDFAQRYAARGHEVLSKWLVGHVCNIDVQLRSR-
31  ------AIIWSDKLSTGLNDIDGDHVNLINLTNEWARDYNKAQIARLREAMLSHFILEEKSISSRASPGRDEHHQRHKSLIDRLDYLVDLFSRGNSQFVYELCIEHILTEDKKVF---
32  ------AIKWRDSLAIGITEIDDQHKKLFEAIDKLFKE----EIKFLEDYTIVHFNDEQEMHKKYNYPERDAHRKIHDNFLNSFSKLKEQFEQEGNKQVVDWLIQHIGKADKAFAA--
33  ------AIKWRDSLAIGITEIDDQHKKLFEAIDKLFKE----EIKFLEDYTIVHFTDEQEMHKKYNYPERDAHRKIHDNFLKSFEKLKEQFENEGNKQVVDWLIQHIGKADKAFAA--
34  ------AIQWDQSLSVGIDEIDNQHKELFSRVNALLDACNQGKIDFLGDYVVTHFGAEERYMQQYGYPDYDAHKAMHDGFIDSFSELKKKFESEGNRTVVDWLINHIGNTDKAL----
35  ------AITWDPTLALGIEEIDAQHEELFRRVDALLERRSA-ELAFLDAYVVEHFGAEEALMQAHRYPGLAAQRAEHAGFAADLAALREELERDGNARVATWLFEHISRSDRAFG---
36  ------AITWKEEYSVGIKTVDEQHKELFARINKLFKD----ALDFLQDYTIFHFNAEQDLMSRAKYPGLEEHKKQHEWFKEQIRSFQEEVQNKGNKVLVDWLINHVTKTDIQYV---
37  ------AITWNSELELGIPVIDSQHRRIVEYINAVYHARNTHSLDELVDYTLSHFAFEENLMEEAGYPFLNAHKKVHRLFARRVGSFQQRVKTGELHVLKAWLINHIKCDDRDYSE--
38  ------AITWRRQLSVGQPAIDDDHKHLIEYLNELDAALAAPRLIKLLEYTKEHFAREERIMQIVHYPKFQEHVAMHRAAVQKVSELSNQFSSDPYKFTADWLVRHIILMDTQLT---
39  ------AIVWKDEYSVGVKVIDDQHKELFRRVNKLFDDVSRGNLDFLNSYVIYHFSAEEQLMAKANYPELESHKNEHEWFKSEILSIREQVEKNGNKLLVSWLINHVTKTDVKFAP--
40  -----AIYSWQDEYSIGIKEIDEQHKELVKMIDELYKR----QLARLVDYCNVHFAAEEALMQSHGYPYYRQHKEIHEKMSARVQALQKCFKAGKSLFIKEWLDKHILGTDQKFAT--
41  -------AKWSPELETGIIWQDLQHKELVEAVKTLYK-----KVAFLYDYTVHHFSMEEKHMEEYGYPDMLQHVDQHQRFIRMLEDFEKERKNSK-----------------------
42  ------ALEWSQALALDLPLMDDTHREFVDLLAAVEAADDAALWQALVEHTEQHFGQEDAWMASTGFASGNCHAMQHKVVLQVMREGATRAVEKGAAELAIWFPQHAQSMDAALA---
43  ------------ALLIDLPEIDVQHEEIFRRIESLKAACFGSGFQSLLDYLKHHFATEEQVALEAA-ADFVDHAKVHRENLHALRRALGEVRHGKLRYAEYWFERHINEEDKPFAAS-
44  ------------ALLVDLPEIDTQHEEIFDRIETLKTACFESSFHSLLELFEQHFATEERLAEEAALD-FVDHTKVHRDTLRILRKALGEVISGALRYSEYWFERHISDDDKLFVA--
45  ------ALQWSEELALDLPQMDHTHQEFVALLQAVESADDAALWEDLVAHTDAHFAQEDRWMQATRFAAGNCHSTQHQVVLKIMREGTDRARQGEAAELAVWFPQHTDAMDASLAQH-
46  ------ALQWSEELALDLPQMDHTHQEFVALLQAVESADDVALWEDLVAHTDAHFAQEDRWMQATRFAAGNCHSTQHQVVLKIMREGTDRARQGEAAELAVWFPQHTDAMDASLAQH-
47  ------ALQWSEGLALDLPLMDETHEEFVALLAAVEQADDEALWRALVAHTTGHFAREDAWMAATRFASGNCHSLQHKVVLQVMGEGTARAEAGEAGELAVWFPQHTQTMDAALA---
48  ------ALQWSEGLALDLPLMDETHEEFVALLAAVEQAGDEALWRALVAHTTEHFAREDAWMAATRFASGNCHSLQHKVVLQVMGEGTARAEAGEAGELAVWFPQHTQTMDAALA---
49  ------ALQWTSALSSGVAFLDAQLEELFRRVDRLLRDRSE-AIAFLQHHVSDHFAAEERLMREVEYRDAARHVAEHLAEAAQIDALAGWLAAEGERFVTAWLQEHVYGTDHALAR--
```

How can having so many sequences change the game in protein bioinformatics?

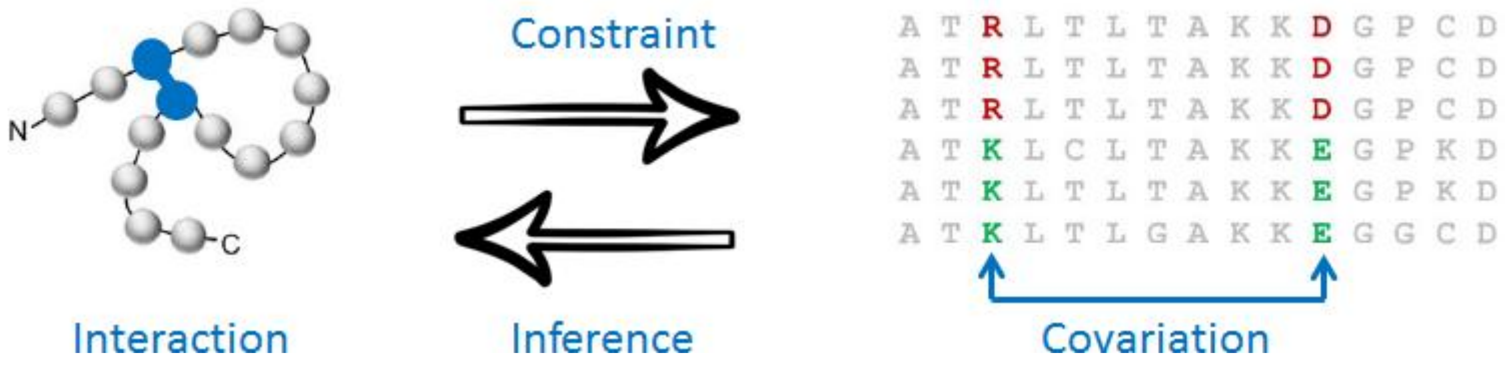
In most molecular evolution applications a common simplifying assumption is made that mutations at one site in a protein occur independently from mutations occurring at other sites

This simplification allows the use of Markovian methods e.g. HMMs and profiles

With massive sequence data sets, however, we can start to consider coevolutionary or epistatic mutational effects



Correlated mutations in Proteins



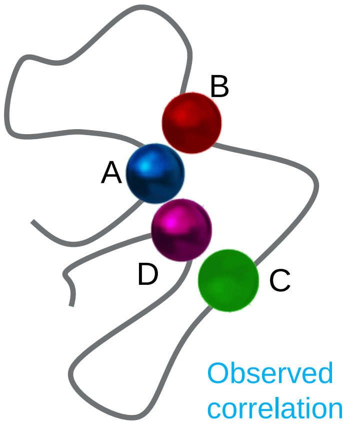
```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
```

- Given enough sequence data, residues in close proximity can be seen to have a tendency to covary, probably to maintain a common microenvironment
- Changes at one site can be compensated for by a mutation in another position. These structural constraints leave a record within the sequence
- By observing patterns of covarying residues in large Multiple Sequence Alignments (MSAs) of homologous protein sequences, we can try to infer this structural information

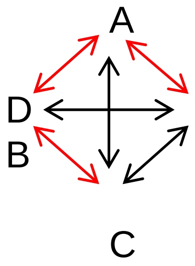
Adapted from Marks, D. S., Hopf, T. A., & Sander, C. (2012). Protein structure prediction from sequence variation. *Nature biotechnology*, *30*(11), 1072-1080.

Contact Correlation Chains

Observed Residue Contacts



Observed Correlations



Observed residue mutations form a chain of correlations even though only 3 physical contacts are made out of the 6 possible pairs

Predicted Contacts

	A		B		C		D
A			🟥		🟦		🟥
B	🟥				🟦		🟦
C	🟦		🟦				🟥
D	🟥		🟦		🟥		

- Not all correlations are the result of direct interactions!
- Direct contact between A and B creates AB correlation
- Direct contact between A and D creates AD correlation
- Direct contact between C and D creates CD correlation

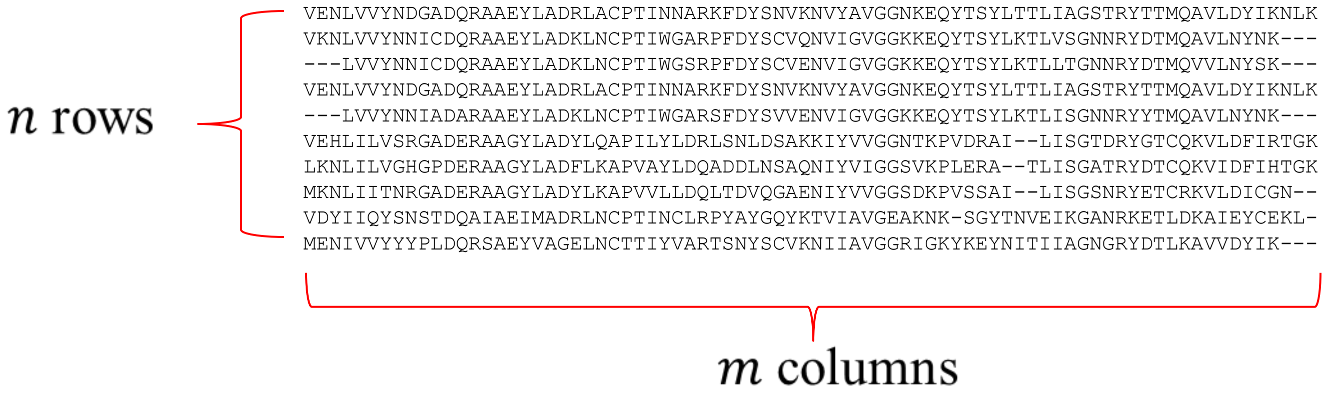
- But - this creates apparent correlation between AC, BC, and BD despite not being in contact
- Analogous to causative/transitive effects in gene network reconstruction

Calculation of Residue Covariation using a Sample Covariance Matrix

We consider an alignment with m columns and n rows, where each row represents a different homologous sequence and each column a set of equivalent amino acids across the evolutionary tree, with gaps considered as an additional amino acid type. We can compute a $21m$ by $21m$ sample covariance matrix as follows:

$$S_{ij}^{ab} = \frac{1}{n}\sum_{k=1}^{n}(x_i^{ak} - \bar{x}_i^a)(x_j^{bk} - \bar{x}_j^b)$$

Where x_i^{ak} is a binary variable ($x \in \{0,1\}$) indicating the presence or absence of amino acid type a at column i in row k and x_j^{bk} the equivalent variable for observing residue type b at column j in row k.



n rows

```
VENLVVYNDGADQRAAEYLADRLACPTINNARKFDYSNVKNVYAVGGNKEQYTSYLTTLIAGSTRYTTMQAVLDYIKNLK
VKNLVVYNNICDQRAAEYLADKLNCPTIWGARPFDYSCVQNVIGVGGKKEQYTSYLKTLVSGNNRYDTMQAVLNYNK---
---LVVYNNICDQRAAEYLADKLNCPTIWGSRPFDYSCVENVIGVGGKKEQYTSYLKTLLTGNNRYDTMQVVLNYSK---
VENLVVYNDGADQRAAEYLADRLACPTINNARKFDYSNVKNVYAVGGNKEQYTSYLTTLIAGSTRYTTMQAVLDYIKNLK
---LVVYNNIADARAAEYLADKLNCPTIWGARSFDYSVVENVIGVGGKKEQYTSYLKTLISGNNRYYTMQAVLNYNK---
VEHLILVSRGADERAAGYLADYLQAPILYLDRLSNLDSAKKIYVVGGNTKPVDRAI--LISGTDRYGTCQKVLDFIRTGK
LKNLILVGHGPDERAAGYLADFLKAPVAYLDQADDLNSAQNIYVIGGSVKPLERA--TLISGATRYDTCQKVIDFIHTGK
MKNLIITNRGADERAAGYLADYLKAPVVLLDQLTDVQGAENIYVVGGSDKPVSSAI--LISGSNRYETCRKVLDICGN--
VDYIIQYSNSTDQAIAEIMADRLNCPTINCLRPYAYGQYKTVIAVGEAKNK-SGYTNVEIKGANRKETLDKAIEYCEKL-
MENIVVYYYPLDQRSAEYVAGELNCTTIYVARTSNYSCVKNIIAVGGRIGKYKEYNITIIAGNGRYDTLKAVVDYIK---
```

m columns

Gaps are considered 21st amino acid type

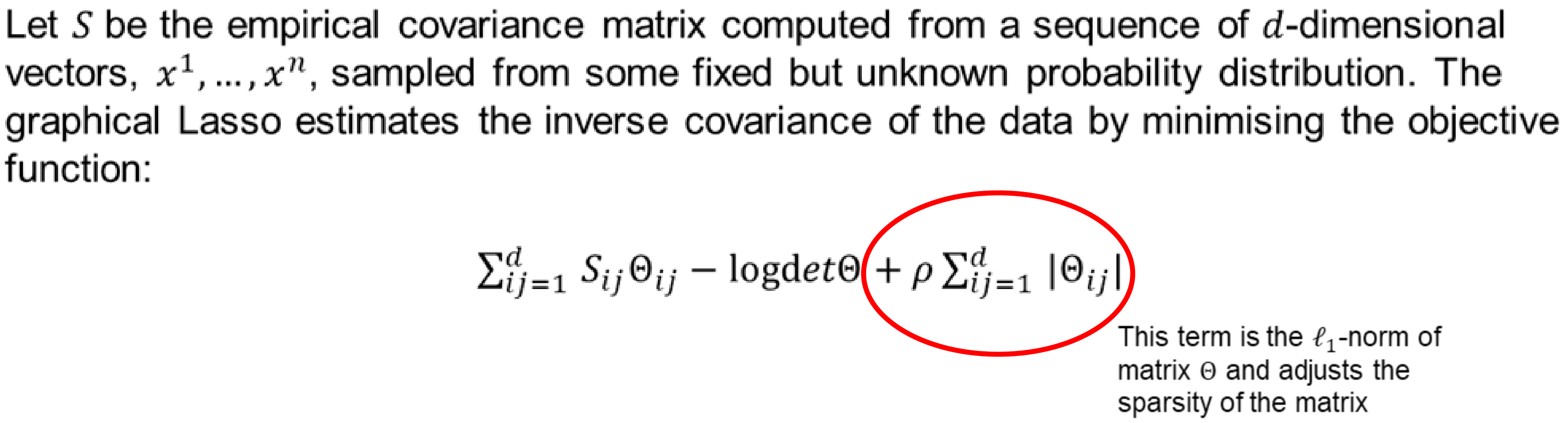
Sparse Inverse Covariance Estimation allows us to solve the coupling problem

The Graphical Lasso attempts to find a sparse inverse covariance matrix by minimization of a specific objective function.

The expected sparsity of the inverse covariance matrix itself provides a powerful self-contained constraint on the obtained solution and thus helps avoid over-fitting.

In general terms, where an inverse covariance estimate is constrained to be sparse, the non-zero terms tend to more accurately relate to correct positive correlations in the true inverse covariance matrix. SICE therefore not only solves the inversion problem but also gives better answers!

The assumption of sparsity in this application is well justified from observations of contacts in known protein structures, where on average only ~3% of all residue pairs are observed to be in direct contact. We actually use this observation to tune the regularisation parameter for each case.

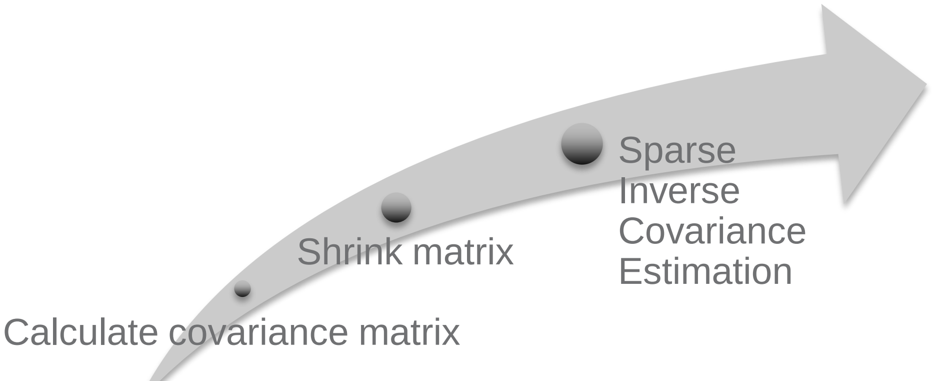


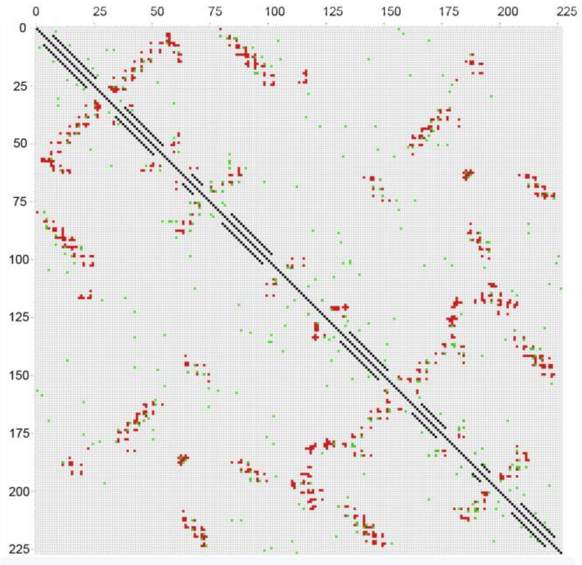
Let S be the empirical covariance matrix computed from a sequence of d-dimensional vectors, $x^1, \ldots, x^n$, sampled from some fixed but unknown probability distribution. The graphical Lasso estimates the inverse covariance of the data by minimising the objective function:

$$\sum_{ij=1}^{d} S_{ij}\Theta_{ij} - \mathrm{logdet}\Theta + \rho \sum_{ij=1}^{d} |\Theta_{ij}|$$

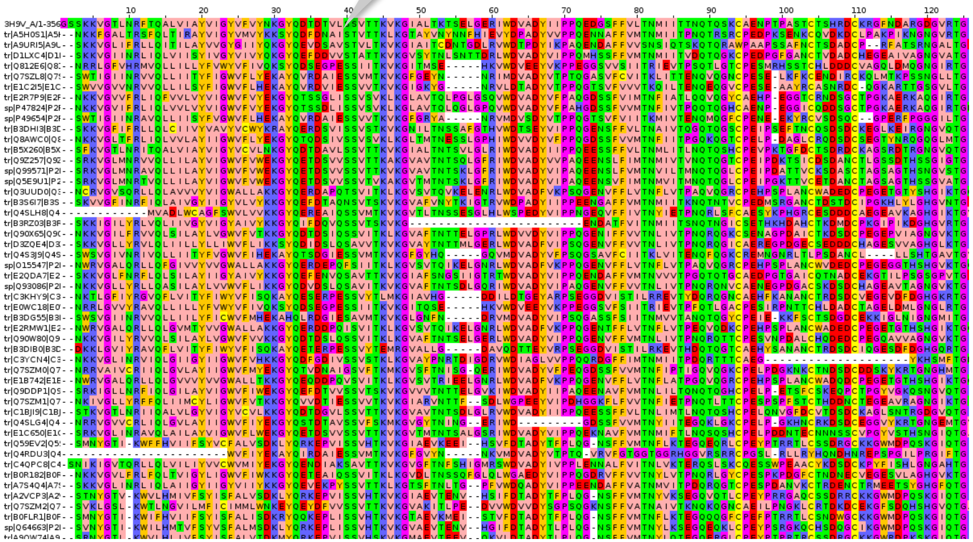
This term is the ℓ_1-norm of matrix Θ and adjusts the sparsity of the matrix

Calculating contact maps using PSICOV


Calculate covariance matrix
Shrink matrix
Sparse Inverse Covariance Estimation



Predicted contact map



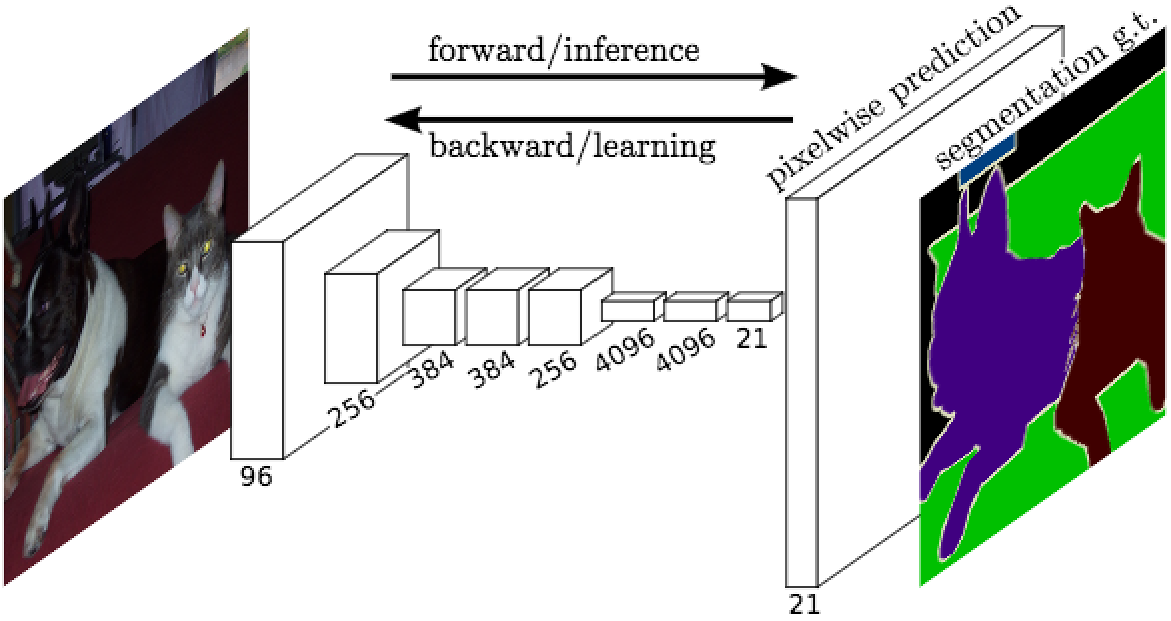
A large multiple sequence alignment is required, >400 sequences

PSICOV – Protein Sparse Inverse COVariance

Jones DT. et al. (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 28:184-190.

THE FRANCIS CRICK INSTITUTE

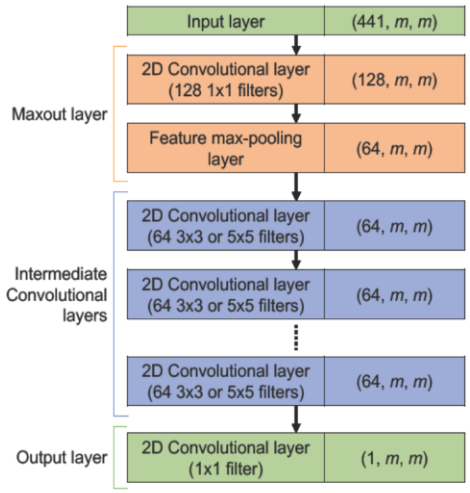
Deep Fully Convolutional Nets can do even better…



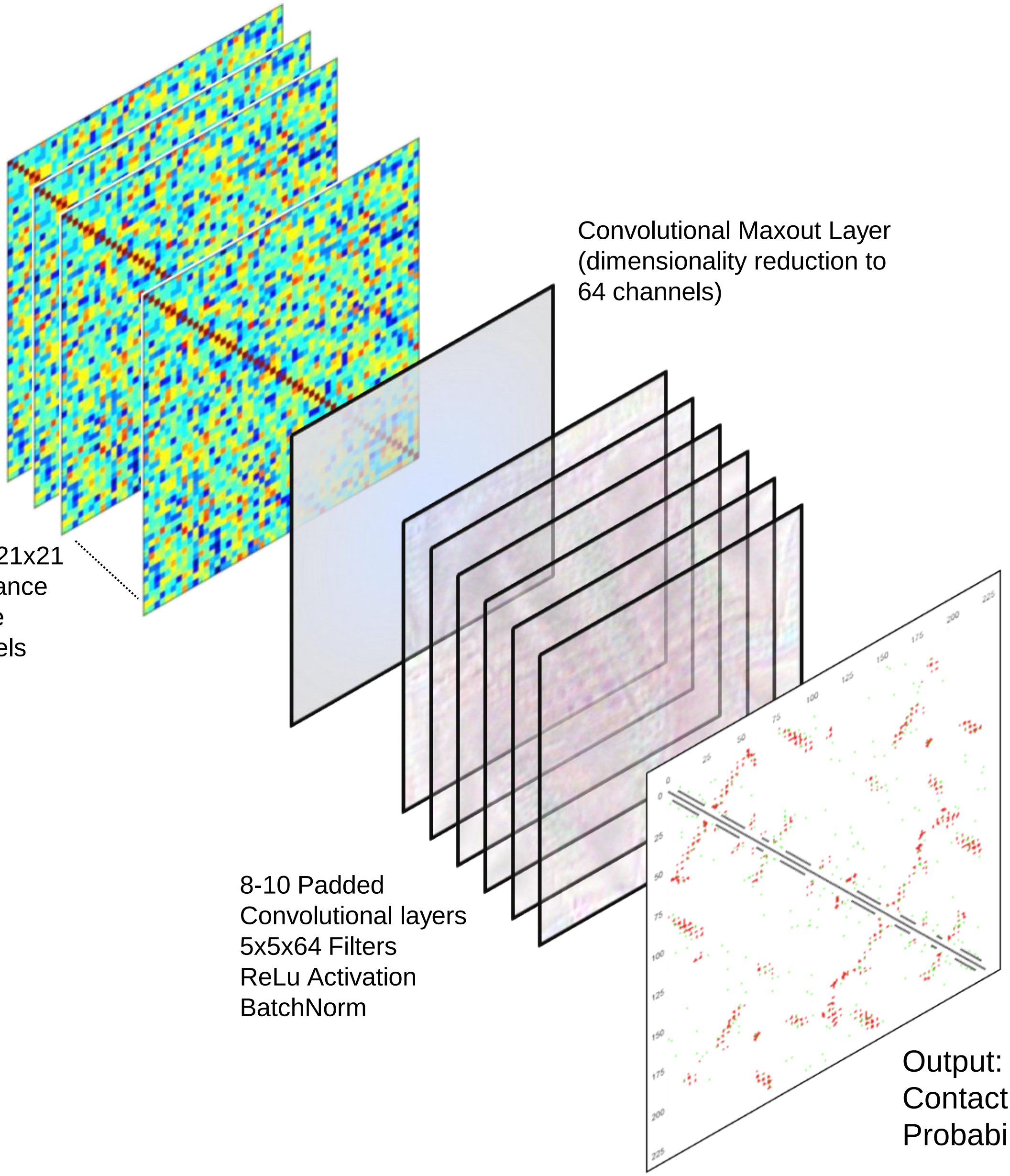
Fully Convolutional Networks for Semantic Segmentation.
Long et al. CVPR2015

DeepCOV: Analysing Residue Covariation using FCNs

THE FRANCIS CRICK INSTITUTE



	Layer	Output
	Input layer	(441, *m*, *m*)
Maxout layer	2D Convolutional layer (128 1x1 filters)	(128, *m*, *m*)
	Feature max-pooling layer	(64, *m*, *m*)
Intermediate Convolutional layers	2D Convolutional layer (64 3x3 or 5x5 filters)	(64, *m*, *m*)
	2D Convolutional layer (64 3x3 or 5x5 filters)	(64, *m*, *m*)
	⋮	
	2D Convolutional layer (64 3x3 or 5x5 filters)	(64, *m*, *m*)
Output layer	2D Convolutional layer (1x1 filter)	(1, *m*, *m*)



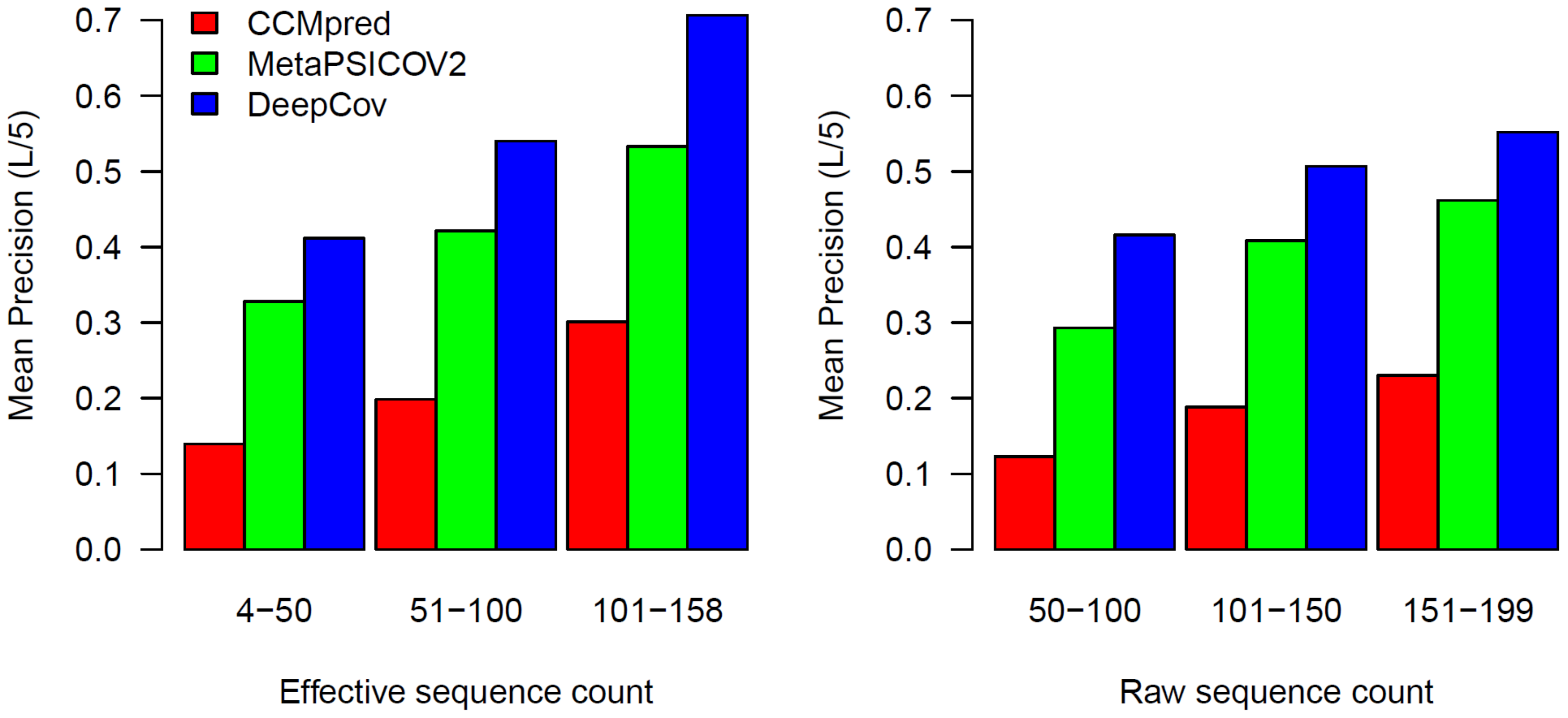
Input: 21x21 covariance feature channels

Convolutional Maxout Layer (dimensionality reduction to 64 channels)

8-10 Padded Convolutional layers
5x5x64 Filters
ReLu Activation
BatchNorm

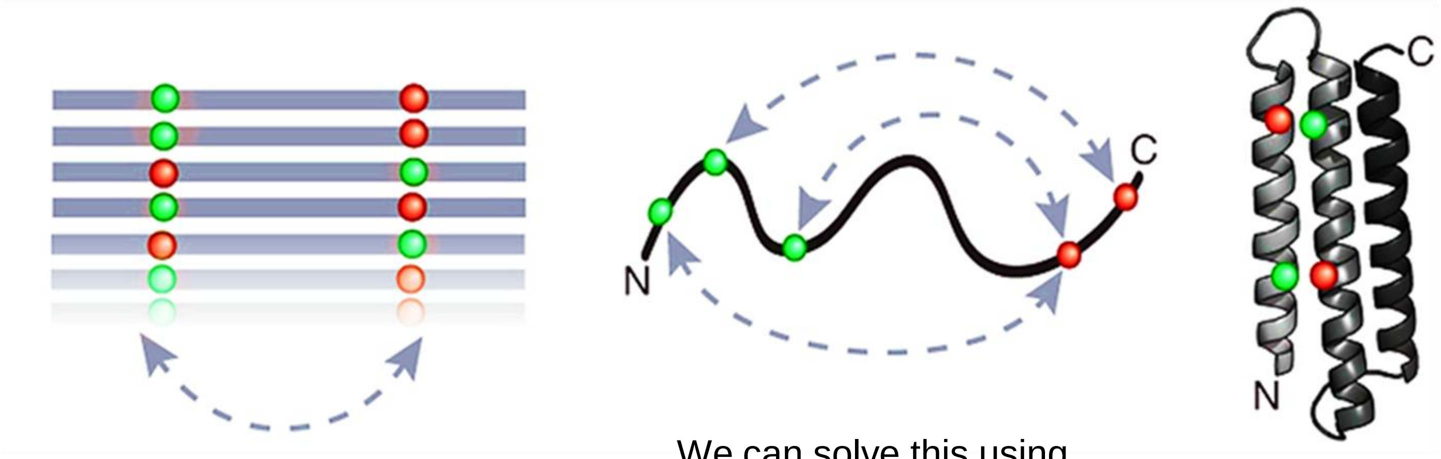
Output: Contact Probability Map

DeepCOV can predict contacts with fewer sequences than previous methods



Predicting the 3-D structure of proteins by co-evolution

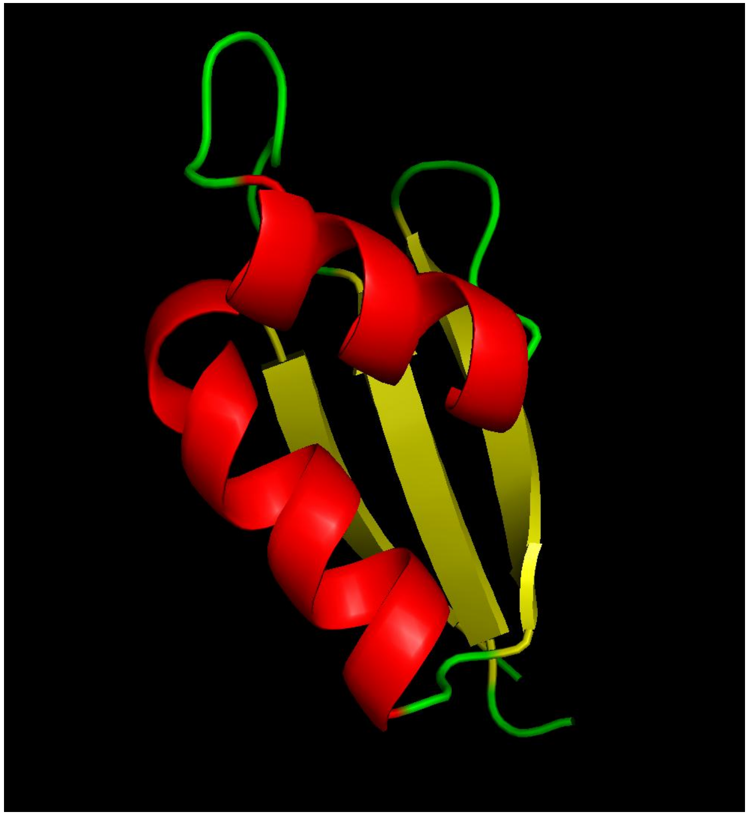
- **We can produce accurate lists of contacting residues from covariation observed in large multiple sequence alignments**
- **If we have an efficient way to project this information into 3-D space whilst satisfying the physicochemical constraints of protein chains then we have everything we need to predict 3-D structure!**



We can solve this using constraint satisfaction methods

Example 1

Ribosomal Protein L30 – 60 amino acids
Total calculation time: 10 min on a single CPU



And another one…

CASP12 Target T0864 – 246 amino acids
Total calculation time: 1 hour on a single CPU

Acknowledgements

Rui Fa
Cen Wan
Domenico Cozzetto
Dan Buchan
Shaun Kandathil