

# Ab Initio Prediction of Protein Interactions with Machine Learning Methods

Wolfgang Lehrach

Institute of Adaptive and Neural Computation (ANC) and  
Biomathematics & Statistics Scotland (BIOSS)

July 2004

## Why are Protein Interactions interesting?

- ▶ Large numbers of proteins with unknown functions - knowing their interaction partners can help to characterise their function.

## Why are Protein Interactions interesting?

- ▶ Large numbers of proteins with unknown functions - knowing their interaction partners can help to characterise their function.
- ▶ When knocking out a pathway in the cell, we should target proteins that are only involved in one interaction pathway, instead of promiscuous proteins

## Why are Protein Interactions interesting?

- ▶ Large numbers of proteins with unknown functions - knowing their interaction partners can help to characterise their function.
- ▶ When knocking out a pathway in the cell, we should target proteins that are only involved in one interaction pathway, instead of promiscuous proteins
- ▶ Ties in with systems biology, helps to infer regulatory networks etc.

## Why are Protein Interactions interesting?

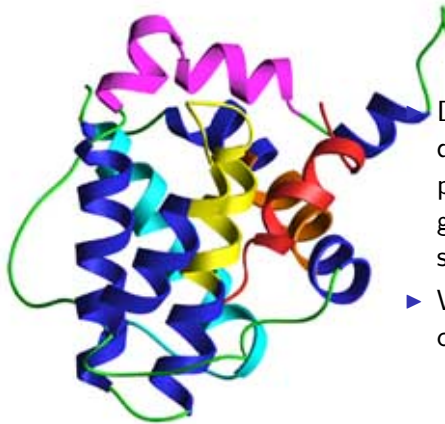
- ▶ Large numbers of proteins with unknown functions - knowing their interaction partners can help to characterise their function.
- ▶ When knocking out a pathway in the cell, we should target proteins that are only involved in one interaction pathway, instead of promiscuous proteins
- ▶ Ties in with systems biology, helps to infer regulatory networks etc.
- ▶ Experimental methods are either large, noisy and uninformative, or small but too expensive to apply to whole organism, so want to complement with computational approaches

# What are Domains and Motifs?



Domains are discrete structural units, defined by their structure, central to protein function whose boundaries can generally be found in the protein sequence.

# What are Domains and Motifs?

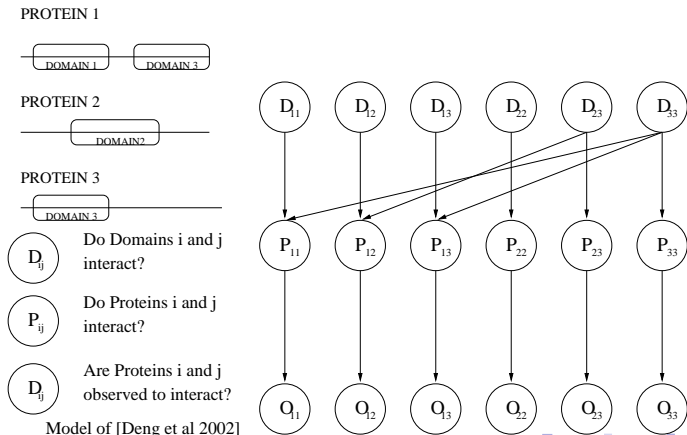


Domains are discrete structural units, defined by their structure, central to protein function whose boundaries can generally be found in the protein sequence.

- ▶ While motifs are simply evolutionary conserved sequences.

## Using Domains/Motifs to predict interactions

- Explains why interactions are occurring. Looking only locally at pairs of domains ignores other possible explanations.



## Adding in sequence information.

- ▶ ... However, this is limited by the power of the domain finding algorithm. Doesn't use protein sequence and ignores that domains evolve over time:

## Adding in sequence information.

- ▶ ... However, this is limited by the power of the domain finding algorithm. Doesn't use protein sequence and ignores that domains evolve over time:
- ▶ Most methods that use sequence information often require mapping all residue sequences to be of equal lengths, done by sub-sampling.

## Adding in sequence information.

- ▶ ... However, this is limited by the power of the domain finding algorithm. Doesn't use protein sequence and ignores that domains evolve over time:
- ▶ Most methods that use sequence information often require mapping all residue sequences to be of equal lengths, done by sub-sampling.
- ▶ Then map these residue sequences to biophysical properties, concatenate together pairs to get a positive pair, and random pairs for a negative pair

## Adding in sequence information.

- ▶ ... However, this is limited by the power of the domain finding algorithm. Doesn't use protein sequence and ignores that domains evolve over time:
- ▶ Most methods that use sequence information often require mapping all residue sequences to be of equal lengths, done by sub-sampling.
- ▶ Then map these residue sequences to biophysical properties, concatenate together pairs to get a positive pair, and random pairs for a negative pair
- ▶ and apply a black box classifier like SVMs. Not that great, as very hard to interpret the output to identify why the interactions are occurring....

## My Model

- ▶ Predicts interactions given only the protein sequences and their interactions

## My Model

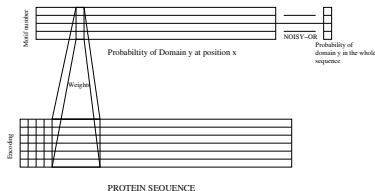
- ▶ Predicts interactions given only the protein sequences and their interactions
- ▶ Central idea is to combine the domain level approaches and local interaction site/domains predictors. Simultaneously take the predicted features for each proteins and use these to predict protein interactions.

## My Model

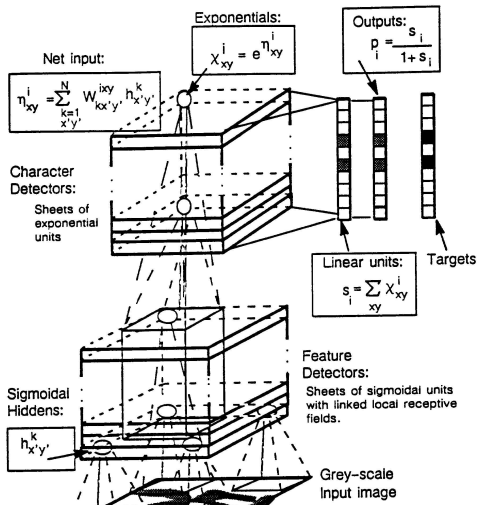
- ▶ Predicts interactions given only the protein sequences and their interactions
- ▶ Central idea is to combine the domain level approaches and local interaction site/domains predictors. Simultaneously take the predicted features for each proteins and use these to predict protein interactions.
- ▶ Inspired by Time Delay Neural Networks...

## My Model

- ▶ Predicts interactions given only the protein sequences and their interactions
- ▶ Central idea is to combine the domain level approaches and local interaction site/domains predictors. Simultaneously take the predicted features for each proteins and use these to predict protein interactions.
- ▶ Inspired by Time Delay Neural Networks...



# A TDNN to Recognise Digits Within Pictures of Zip-codes



## Model setup

Predicting interactions given which domains are present:

$$P(P_{ij}) = 1 - \prod_{k,l} (1 - P(D_{ki} = 1) P(D_{lj} = 1) \lambda_{kl}) \quad (1)$$

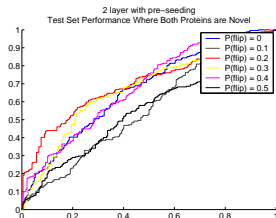
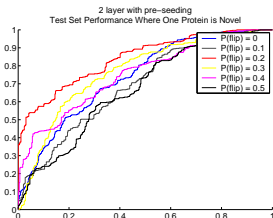
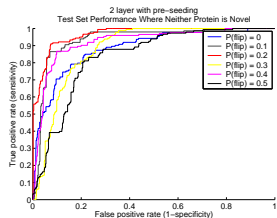
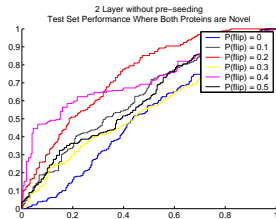
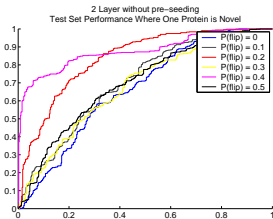
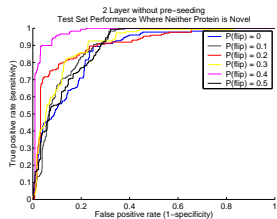
Summarising which motifs are present on the whole protein:

$$P(D_{li} = 1) = 1 - \prod_o (1 - P(F_{lo}^i = 1)) \quad (2)$$

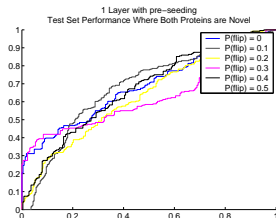
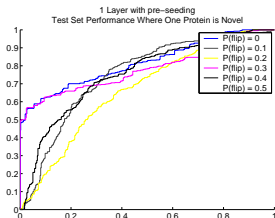
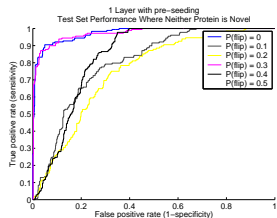
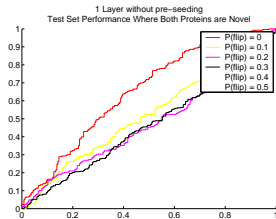
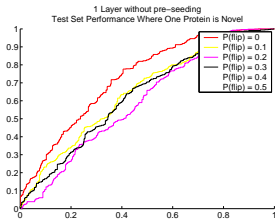
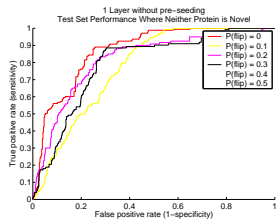
Detecting the presence of a motif in a given location:

$$P(F_{lo}^j | S^j) = \text{logit} \sum_{r=1}^r \sum_{m=1}^{M^l} W_{r,m}^l S_{r,o+m-1} + T^l \quad (3)$$

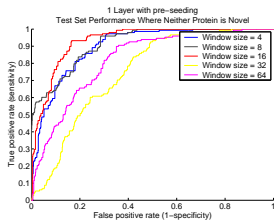
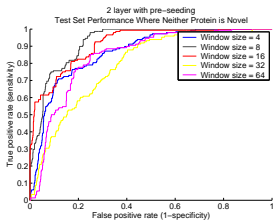
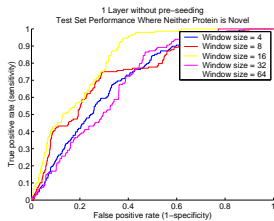
# Effect of varying the noise levels on artificial data - 1 Layer network



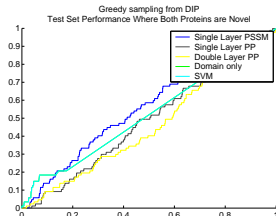
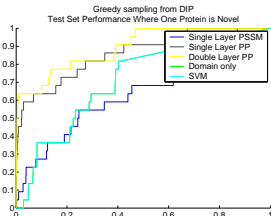
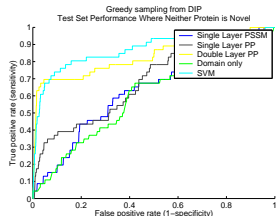
# Effect of varying the noise levels on artificial data - 2 Layer network



# Effect of varying the window size length

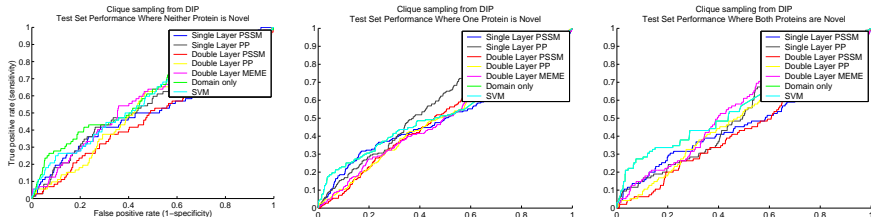


## Clique sampling



- ▶ Performance on a simple subset of DIP CORE Yeast - performance increases significantly with larger subsets as more information is known about each protein.

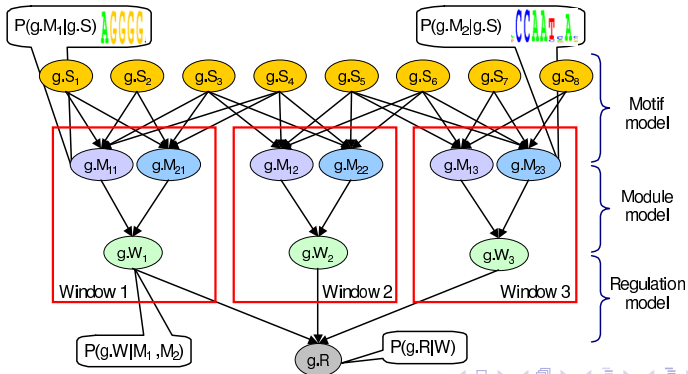
## Greedy sampling



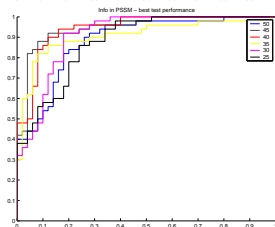
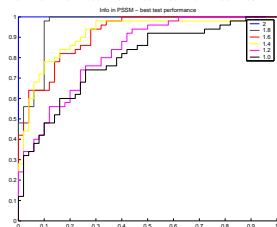
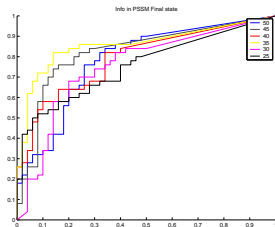
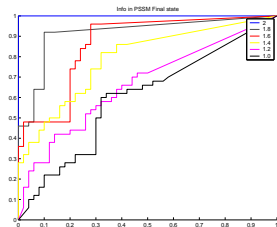
- ▶ Performance on a simple subset of DIP CORE Yeast, based around cliques - performance increases significantly with larger subsets as more information is known about each protein.

## A similar model to detect CRMs

- ▶ [Segal et al 2004] build a discriminative model that simultaneously finds the modules and uses them to predict gene regulation, roughly same principle. Uses EM + conjgrad descent for max step.



## Applying our model to identifying CMS:



## Identifying the motifs to which SH3 binds

- ▶ [Reiss et al 2004] presented at ISMB 04 a method to identify protein motifs which the SH2 domain binds to.

## Identifying the motifs to which SH3 binds

- ▶ [Reiss et al 2004] presented at ISMB 04 a method to identify protein motifs which the SH2 domain binds to.
- ▶ Constructive model that uses Gibbs sampling to find motifs that are similar to a global motif but dis-similar to each other - incorporates this using the prior information.

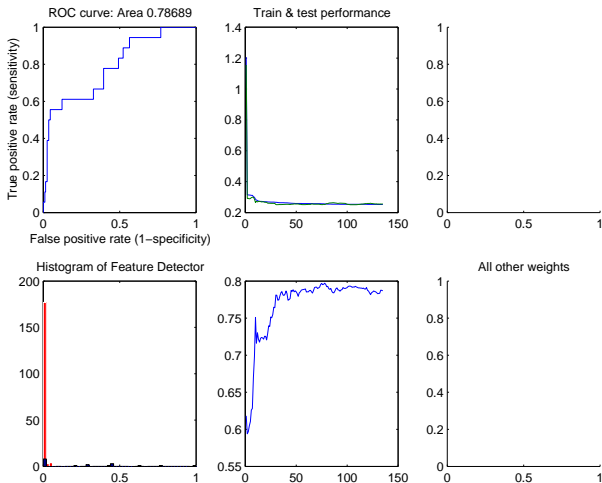
## Identifying the motifs to which SH3 binds

- ▶ [Reiss et al 2004] presented at ISMB 04 a method to identify protein motifs which the SH2 domain binds to.
- ▶ Constructive model that uses Gibbs sampling to find motifs that are similar to a global motif but dis-similar to each other - incorporates this using the prior information.
- ▶ Data is from [Tong et al 2002] and consists of yeast two hybrid of SH3 domains versus all proteins, which contains explicit negative interaction. Will be very interesting test case.

## Identifying the motifs to which SH3 binds

- ▶ [Reiss et al 2004] presented at ISMB 04 a method to identify protein motifs which the SH2 domain binds to.
- ▶ Constructive model that uses Gibbs sampling to find motifs that are similar to a global motif but dis-similar to each other - incorporates this using the prior information.
- ▶ Data is from [Tong et al 2002] and consists of yeast two hybrid of SH3 domains versus all proteins, which contains explicit negative interaction. Will be very interesting test case.
- ▶ But depends on various hand-tweaked parameters - goes horribly wrong if

# Preliminary Results on Tong



## Current Possible Model Improvements

- ▶ More emphasis a few discrete motifs present in sequence, rather than mixtures of lots of sequence fragments which results in recognising proteins instead of relevant motifs. For instance, problems when running on a data set containing a single, all it needs is one detector to recognise any motif on the central protein and one to always fire.

## Current Possible Model Improvements

- ▶ More emphasis a few discrete motifs present in sequence, rather than mixtures of lots of sequence fragments which results in recognising proteins instead of relevant motifs. For instance, problems when running on a data set containing a single, all it needs is one detector to recognise any motif on the central protein and one to always fire.
  - ▶ Add in our expectations into the initialisation stage

## Current Possible Model Improvements

- ▶ More emphasis a few discrete motifs present in sequence, rather than mixtures of lots of sequence fragments which results in recognising proteins instead of relevant motifs. For instance, problems when running on a data set containing a single, all it needs is one detector to recognise any motif on the central protein and one to always fire.
  - ▶ Add in our expectations into the initialisation stage
  - ▶ Ensure that we never test on a single clique

## Current Possible Model Improvements

- ▶ More emphasis a few discrete motifs present in sequence, rather than mixtures of lots of sequence fragments which results in recognising proteins instead of relevant motifs. For instance, problems when running on a data set containing a single, all it needs is one detector to recognise any motif on the central protein and one to always fire.
  - ▶ Add in our expectations into the initialisation stage
  - ▶ Ensure that we never test on a single clique
  - ▶ Add prior information about what a motif looks like/how often it occurs

## Current Possible Model Improvements

- ▶ More emphasis a few discrete motifs present in sequence, rather than mixtures of lots of sequence fragments which results in recognising proteins instead of relevant motifs. For instance, problems when running on a data set containing a single, all it needs is one detector to recognise any motif on the central protein and one to always fire.
  - ▶ Add in our expectations into the initialisation stage
  - ▶ Ensure that we never test on a single clique
  - ▶ Add prior information about what a motif looks like/how often it occurs
  - ▶ Constructive model - implicitly contains more information about how a motif relates to the background

## Current Possible Model Improvements

- ▶ More emphasis a few discrete motifs present in sequence, rather than mixtures of lots of sequence fragments which results in recognising proteins instead of relevant motifs. For instance, problems when running on a data set containing a single, all it needs is one detector to recognise any motif on the central protein and one to always fire.
  - ▶ Add in our expectations into the initialisation stage
  - ▶ Ensure that we never test on a single clique
  - ▶ Add prior information about what a motif looks like/how often it occurs
  - ▶ Constructive model - implicitly contains more information about how a motif relates to the background
- ▶ Cleaning up found motifs.

## Current Possible Model Improvements

- ▶ More emphasis a few discrete motifs present in sequence, rather than mixtures of lots of sequence fragments which results in recognising proteins instead of relevant motifs. For instance, problems when running on a data set containing a single, all it needs is one detector to recognise any motif on the central protein and one to always fire.
  - ▶ Add in our expectations into the initialisation stage
  - ▶ Ensure that we never test on a single clique
  - ▶ Add prior information about what a motif looks like/how often it occurs
  - ▶ Constructive model - implicitly contains more information about how a motif relates to the background
- ▶ Cleaning up found motifs.
- ▶ Redundancy between motifs.

## Current Possible Model Improvements

- ▶ More emphasis a few discrete motifs present in sequence, rather than mixtures of lots of sequence fragments which results in recognising proteins instead of relevant motifs. For instance, problems when running on a data set containing a single, all it needs is one detector to recognise any motif on the central protein and one to always fire.
  - ▶ Add in our expectations into the initialisation stage
  - ▶ Ensure that we never test on a single clique
  - ▶ Add prior information about what a motif looks like/how often it occurs
  - ▶ Constructive model - implicitly contains more information about how a motif relates to the background
- ▶ Cleaning up found motifs.
- ▶ Redundancy between motifs.
- ▶ Convergence issues

## Comments + Future Work

- ▶ We have derived but not yet implemented a Bayesian soft weight sharing scheme that incorporates extra biological knowledge

## Comments + Future Work

- ▶ We have derived but not yet implemented a Bayesian soft weight sharing scheme that incorporates extra biological knowledge
- ▶ Can easily be applied to both DNA and Protein sequence and predicting interactions between them. We are currently looking into available datasets.