A Protocol to Detect Local Affinities Involved in Proteins Distant Interactions

Christophe N. Magnan, Cécile Capponi, François Denis

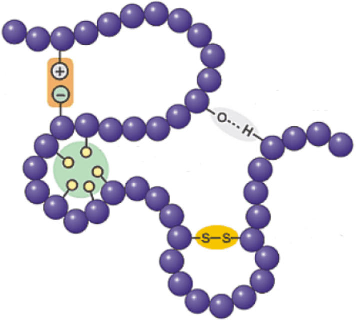
LIF, CNRS, France

UNIVERSITÉ DE PROVENCE
Aix - Marseille 1

The IEEE International conference on Bioinformatics and Biomedicine 2007
BIBM 2007
Silicon Valley, USA
November 2 - 4, 2007

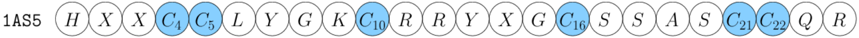
Proteins distant interactions - examples

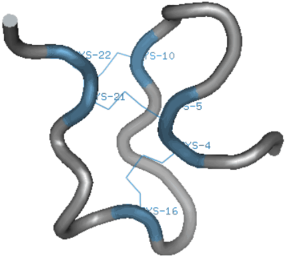
Ionic links, hydrogen bonds, hydrophilic interactions, salt bridges, disulfide bridges, Van Der Waals forces, ...

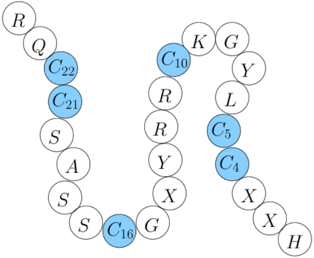


3D structure constrained/stabilized by these interactions

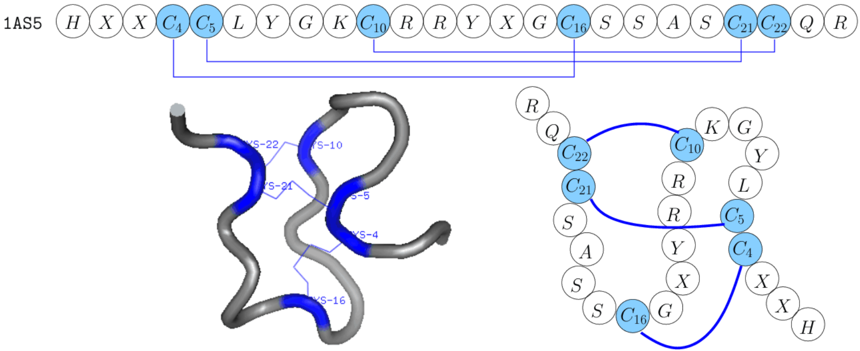
Prediction of disulfide bridges: a two-stage process


1AS5
H X X C_4 C_5 L Y G K C_{10} R R Y X G C_{16} S S A S C_{21} C_{22} Q R

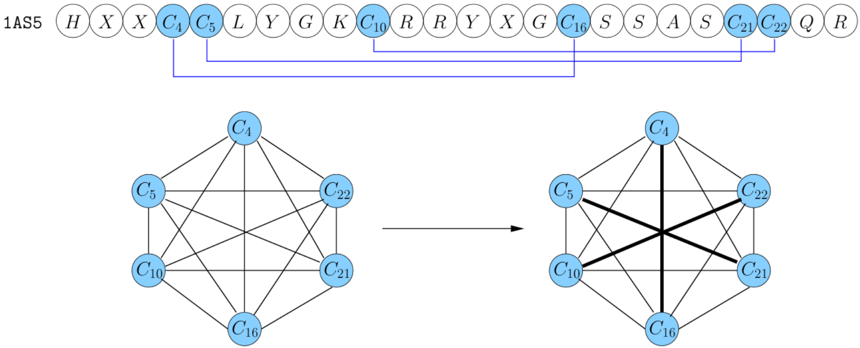



R Q C_{22} C_{21} S A S S C_{16} G X Y R R C_{10} K G Y L C_5 C_4 X X H

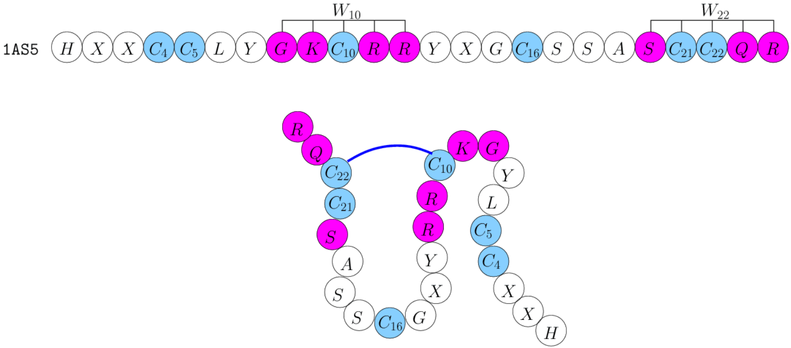
Prediction of disulfide bridges: a two-stage process


1AS5
H X X C_4 C_5 L Y G K C_{10} R R Y X G C_{16} S S A S C_{21} C_{22} Q R
YS-22
YS-10
YS-21
S-5
YS-4
YS-16
R Q C_{22} C_{21} S A S S C_{16} G X Y R R C_{10} K G Y L C_5 C_4 X X H

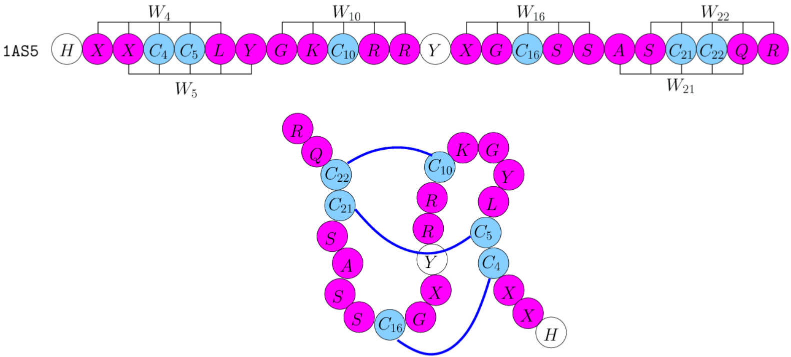
Prediction of disulfide bridges: a two-stage process



Local environments of bonded amino-acids


W_{10}
W_{22}
1AS5
H X X C_4 C_5 L Y G K C_{10} R R Y X G C_{16} S S A S C_{21} C_{22} Q R
R
Q
C_{22}
C_{21}
S
A
S
S
C_{16}
G
X
Y
R
R
C_{10}
K
G
Y
L
C_5
C_4
X
X
H

Local environments of bonded amino-acids


W_4
W_{10}
W_{16}
W_{22}
1AS5
H X X C_4 C_5 L Y G K C_{10} R R Y X G C_{16} S S A S C_{21} C_{22} Q R
W_5
W_{21}

Is there information carried by local environments involved in the formation of bonds such as disulfide bridges?

Local information?

Are the local environments involved in interactions?

- β-sheets: there is local information
- Disulfide/salt bridges: no biological evidence
- Some biologists and biochemists skeptical

Local information?

Are the local environments involved in interactions?

- β-sheets: there is local information
- Disulfide/salt bridges: no biological evidence
- Some biologists and biochemists skeptical
- Always used to predict disulfide bridges

Local information?

Are the local environments involved in interactions?

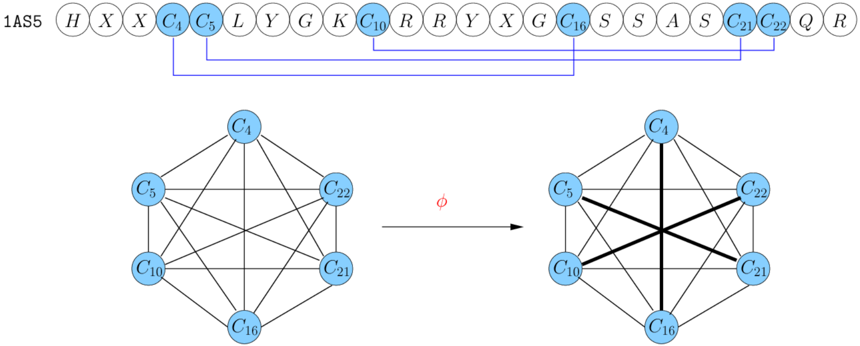
- β-sheets: there is local information
- Disulfide/salt bridges: no biological evidence
- Some biologists and biochemists skeptical
- Always used to predict disulfide bridges

- Is it possible to detect such information?
- Is it possible to show that there exists an **affinity** between local environments of bonded residues involved in the pairing of these residues?

Model

- Σ: the set of 20 amino-acids,
- $\mathcal{P} \subset \Sigma^*$: proteins containing an even number of amino-acids involved in bridges
- $\mathcal{P}_l \subset \mathcal{P}$, proteins with $2l$ amino-acids involved in bridges
- ϕ, a function which associates the correct connectivity to a protein in $\mathcal{P}$

Model


1AS5
H X X C_4 C_5 L Y G K C_{10} R R Y X G C_{16} S S A S C_{21} C_{22} Q R
C_4
C_5
C_{22}
C_{10}
C_{21}
C_{16}
ϕ
C_4
C_5
C_{22}
C_{10}
C_{21}
C_{16}

The prediction of interactions between amino-acids amounts to approximating ϕ with the highest precision

Model

Local environments: segments centered on bonded amino-acids of size $2r+1$ are considered.

- P a distribution over $\mathcal{P}$
- $\Omega_r = \Sigma^{2r+1}$ the set of proteins segments of size $2r+1$
- For $w, w' \in \Omega_r$ let:
  - $P(w)$ the probability that w is a local environment
  - $P(B(w,w')|w,w',I)$ the probability that w and w' are bonded knowing that there are distinct local environments of amino-acids involved in interactions into a protein $p \in \mathcal{P}_I$

Local information?

Let p be a protein with l bridges ($2l$ involved amino-acids).

$$P(B(w, w')|w, w', l) = \frac{1}{2l-1} \Leftrightarrow \text{No local information for pairing amino acids}$$

- a probabilistic way to determine if the local context of bonded residues is involved into the formation of the bridges

- but, estimating directly these probabilities is impossible:

  $r = 3 \rightarrow |\{(w, w'), w, w' \in \Omega_r\}| = 20^{12} \simeq 4 \cdot 10^{15}$, while only few hundreds examples are available in databases!

An affinity function g

The solution we propose:

To suppose the existence of an affinity function $g : \Omega_r^2 \rightarrow Y$ ($|Y|$ small) such that:

$$g(w_1, w_2) = g(w_1', w_2') \Rightarrow P(B(w_1, w_2)|w_1, w_2, l) \simeq P(B(w_1', w_2')|w_1', w_2', l)$$

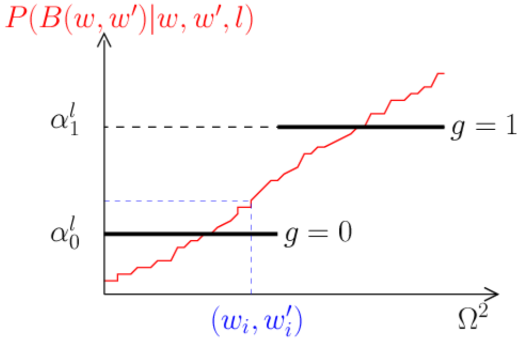
and

$$y < y' \Rightarrow P(B(w_1, w_2)|g(w_1, w_2) = y) < P(B(w_1', w_2')|g(w_1', w_2') = y')$$

A simple case: $Y = \{0, 1\}$

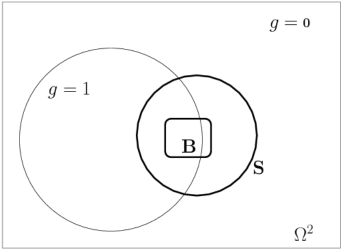
With $Y = \{0, 1\}$, pairs of local environments are partitioned into two classes, corresponding to two affinity levels and:

$$P(B(w,w')|w,w',l) \simeq P(B(w,w')|g(w,w'),l) = \begin{cases} \alpha_1^l \textit{ if } g(w,w') = 1 \\ \alpha_0^l \textit{ if } g(w,w') = 0 \end{cases}$$



Observations as indirect information on g

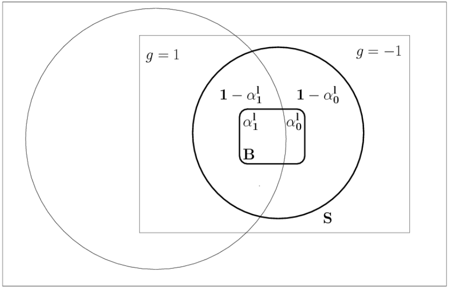
The observed classes (bonded or non-bonded) of examples issued from experiments do not carry direct information about g.



Observed pairs as noisy examples of g

The pairs such that

- g=1 correspond to observing a bridge with noise $\eta^+ = 1 - \alpha_1^l$
- g=0 correspond to non-bonded pairs with noise $\eta^- = \alpha_0^l$



- generalization of the *uniform classification noise* ($\eta^+ = \eta^-$)
- referred to as *class-conditional classification noise* (CCCN)

Setting up the protocol to learn g

If a local information exists

If it can be represented by a function learnable under CCCN

then we should be able to detect, extract and evaluate it

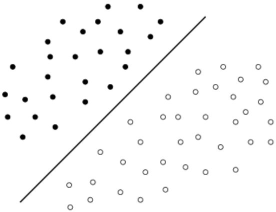
assuming that we have access to a sufficient number of examples

What can we learn under CCCN?

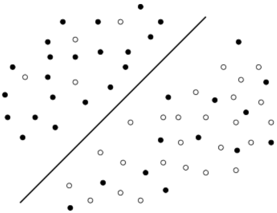
- some theoretical results
- they can not be used in practice
- methods such as Soft-margins SVM cannot handle data corrupted by CCCN
- New methods have to be created

Perceptron CCCN

- we propose an algorithm to learn *linear threshold functions* from examples corrupted by CCCN
- a generalization of the Perceptron algorithm



No noise



CCCN

Datasets

- 1 dataset of proteins featuring salt bridges: G3D
    - 1836 internal salt bridges in 570 proteins
    - created from PDB by Christophe Geourjon (IBCP, Lyon, France) in 2005
- 1 dataset of proteins featuring disulfide bridges: SPX
    - 1676 internal disulfide bridges within 567 proteins
    - created from Swiss-Prot by Jianlin Cheng and Pierre Baldi (Irvine, California) in 2005
- proteins containing from 2 to 5 bonds

Coding of local environments pairs

For a protein containing l bridges:

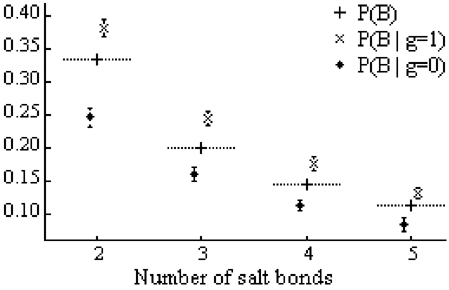
- $l(2l-1)$ pairs of local environments
- radius $r = 6$ ($|w| = 13$)
- each local environments pair (w, w') is described as follows:
    - 169 amino-acids pairs (a_i, a_j), with $a_i \in w$ and $a_j \in w'$ ($i, j \in \{1, ..., 13\}$)
    - (w, w') is modeled with a vector of $\mathbb{R}^m$ with:
        - m is the number of ordered pairs of amino-acids in Σ ($m = 231$)
        - each coordinate is the number of time the corresponding pair is observed in (w, w')

Experiments and studied criteria

- we launch 5 10-fold cross-validations for both kinds of bonds
- two criteria are studied:
    - $P(B|g=1)$, the probability to observe a bond knowing that the pair is predicted to have a high level of affinity
    - $P(B|g=0)$, the probability to observe a bond knowing that the pair is predicted to have a low level of affinity

Salt bridges

A clear signal is detected:
$\forall l \in \{2,3,4,5\},\ P(B|g=1,l) > P(B|g=0,l)$



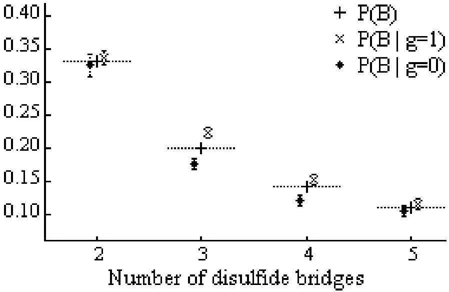
Salt bridges

The detected affinities might be explained either by

- the ionic nature of salt bridges
- the hydrophilic property of many residues around salt bridges

Disulfide bridges

Results are not as clear as expected:
$\forall l \in \{2,3,4,5\},\ P(B|g=1,l) \simeq P(B|g=0,l)$



Disulfide bridges

These results may be explained by several independent reasons:

- Biology reality: there might be no local information that would guide the formation of disulfide bridges
- Learning a function in an unsuitable function class: the function g that we try to learn might be not representable by a linear threshold function.
- ...

This work give us no hint on which assumption is the most probable

Conclusions

- a machine-learning based protocol to answer the question of the presence of local affinities
- independent from the contact
- results on salt bridges validate this protocol
- disulfide bridges remain an open question