DBDB : a Disulfide Bridge DataBase for the predictive analysis of cysteine residues involved in disulfide bridges

Emmanuel Jaspard[1] Gilles Hunault[2] Jean-Michel Richer[3]

[1] Laboratoire PMS UMR A 1191, Université d'Angers, 2 Boulevard Lavoisier, 49045 Angers Cedex 01
emmanuel.jaspard@univ-angers.fr
[2] Université d'Angers, 2 Boulevard Lavoisier, 49045 Angers Cedex 01
gilles.hunault@univ-angers.fr
[3] LERIA, Université d'Angers, 2 Boulevard Lavoisier, 49045 Angers Cedex 01
jean-michel.richer@univ-angers.fr

Abstract: *We present a database of structured, verified and useful information about cysteine residues involved in disulfide bridges. The database has been filled with proteins of a subset of the PDB database called* select 25[4]. *A rough analysis of the proteins contained in the database shows that it is possible to identify cysteine residues which are involved or not in a disulfide bond from the amino acid sequence that surrounds the cysteine. The goal of this database is to serve as a reference for the evaluation of disulfide bridge prediction softwares and will help us develop our own prediction software.*

Keywords: Disulfide bridge prediction, database, bioinformatics

1. Biological context

Proteins contain cysteine residues that can be oxidized to form a covalent bond called a disulfide bridge. Past experiments [2] showed that disulfide bridges can increase the thermodynamic stability of the native structure of proteins by reducing the number of unfolded conformations. Therefore, an exact prediction of disulfide connectivity can strongly reduce the conformational search space and increase the accuracy in protein structure prediction. Based on an analysis of a part of the Protein Data Bank (PDB) [3], the select 25 subset, we found that about 25% of the proteins of this subset contain disulfide bridges.

Cysteine residues are considered as free if they are not involved in a disulfide bridge. Otherwise they can be implied in an intra or inter bond if they belong to the same polypeptidic chain or to two different chains, respectively. Note that the PDB select 25 contains only 3% of inter bridges[5].

2. Presentation and purpose of the Disulfide Bridge DataBase (DBDB)

Up to now it has not been possible to assert if a given cysteine is involved in a disulfide bridge, unless the three dimensional structure of the protein is known.

[4] http://homepages.fh-giessen.de/~hg12640/pdbselect/
[5] This analysis is based on the data downloaded from ftp://ftp.rcsb.org/pub/pdb/data/biounit/coordinates/all

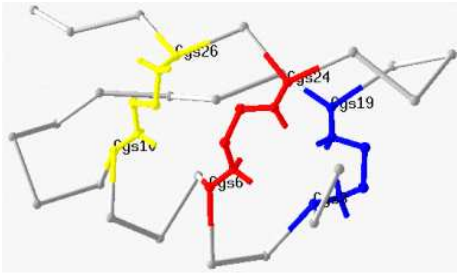


Figure 1. The three intra disulfide bridges of the scorpion toxin peptide P01 (PDB 1ACW). The sequence is VSC^{3}EDC^{6}PEHC^{10}STQKAQAKC^{19}DNDKC^{24}VC^{26}EPI.

Nevertheless much work has been devoted to the prediction of disulfide bonds [4,1,5,6]. The main drawback of these approaches is that they rely on sets of proteins, called by extension databases which are not organized as genuine databases, but which are a list of PDB identifiers. Some other researchers have developed specialized databases [8,7] for the study of disulfide bridges. Our goal is to provide the bioinformatics community with a database of selected proteins issued from the PDB, that we have called DBDB for Disulfide Bridge DataBase, which has two main purposes :

- the first one is to provide structured, verified and useful information on the number of intra and/or inter bonds of a protein and of the positions of cysteine residues involved or not in disulfide bridge; the taxonomy and classification information of proteins of the DBDB allows to focus on specific sets of proteins in order to study functions and/or species,
- the second one is to serve has a benchmark set specifically designed for the evaluation and comparison of cysteine bridges prediction softwares [1,9].

The database is currently available at the following address :

`http://www.info.univ-angers.fr/pub/richer/rec/bio/dbdb`

and contains up to now 356 proteins, 84123 amino acids among which 2042 cysteines, organized as follows (see table 1) :

The first set of proteins (*Working set*) contains proteins with intra and/or inter bonds. The second set contains proteins with no bonds and serves as a *Control set*. This discrimination enables us to define four kinds of environments and their related sets for cysteine residues (see fig. 2) :

- T, the set of cysteines which belong to proteins with no bond
- for cysteines which belong to proteins
  - F, the set of free cysteines, not involved in a bond
  - I_a, the set of cysteines involved in an intra bond
  - I_e, the set of cysteines involved in an inter bond

Each entry in the Disulfide Bridge DataBase contains the following type of information which are mostly extracted from the PDB files :

- the PDB identifier of the protein,

	Working Set	*Control set*
Containing	proteins with intra/inter bonds	proteins without bonds
Number of proteins	300	56
Number of chains	349	60
Number of intra bonds	654	*not applicable*
Number of inter bonds	42	*not applicable*
Number of free cysteines	359	291

Cysteines		*PolyPeptidic chains*	
Total	1751	Total	356
intra bond	1308	with no bond	(Control set) 56
inter bond	84	with intra bond	278
free	359	with inter bond	33
		with intra and inter bonds	11

Table 1. Information currently available in the Disulfide Bridge DataBase

- the classification of the protein such as hydrolase or toxin,
- for enzymes, the E.C. number,
- the taxonomy of the organism like *Homo sapiens*, *Bos taurus*,
- the positions of cysteines (for intra and inter bonds); these positions are validated, i.e., they were manually adjusted when discrepancies between the Fasta sequence and the positions given by the PDB files were found,
- bibliographic information.

The DBDB allows the user to view cysteine residues directly positioned in the Fasta sequence of each chain that composes the protein (see fig 6).

3. Analysis of cysteine environments

In the following we will compare the evolution of frequencies of occurrence of amino acids that belong to the environment of a cysteine. This analysis is based on the amino acid sequence of proteins.

3.1 Amino acid frequencies

The whole database contains more than 80000 residues. The most present amino acids are L, G, and A, while the less present are respectively W, M, H and C. The diagram on figure 3 represents the frequencies of occurrence of each amino acid in the database.

3.2 Statistical analysis of amino acid frequencies

A first analysis of the DBDB was performed in order to determine the influence of amino acids that appear around all cysteines as they are considered to influence the formation of disulfide bridges. If a residue is more

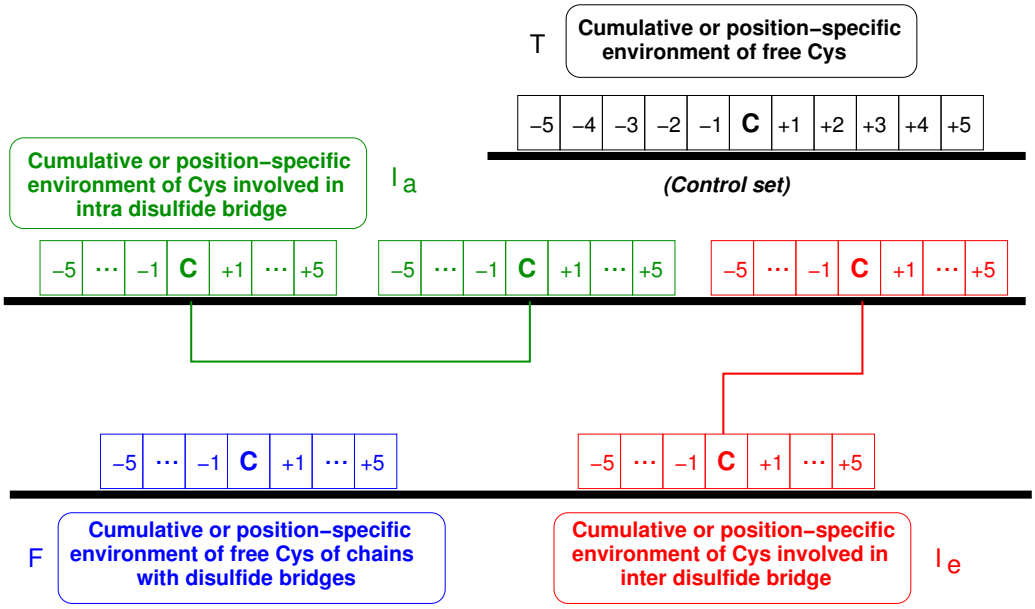


Figure 2. The four different kinds of cysteine environments in proteins

(or less) present than it should be in one kind of environment, then we might infer that it has some influence on the environment and might play some key role.

We first made two separate analysis for the Working and Control sets. For the Working set we studied the three subsets of segments for the different environments : F, I_a and I_e. For the Control set we can only focus on the free cysteine set T.

For each segment of 11 residues (see [10] for the choice of the segment size) centered on a cysteine we computed the difference between the frequency of occurrence of an amino acid near the Cys residue and its global frequency of occurrence over the whole database. We then performed

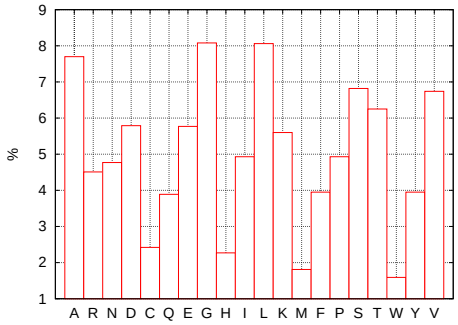


Figure 3. Percentage of occurrence of each amino acid in DBDB

- a *cumulative* analysis of all segments for each environment taking into account residues $X_{-n} \dots X_{-1}$ C $X_{+1} \dots X_{+n}$, for a given distance $n \in [1..5]$ in order to study the global influence of amino acids,
- a *position-specific* analysis only taking into account residues X_{-n} and X_{+n} in order to determine the local evolution of the frequencies of amino acids.

From a general point of view we can say that :

- the four kinds of environments are different and show different variations of frequencies of amino acids around cysteine residues, except for D which is not influent (see graphs 2,3,4 of figure 4).
- the inter bond environment I_e shows the most important variations of the frequencies for a great number of amino acids : V, A, F, R, S, W, N, Q and G. In particular G is overabundant when it is very close to the cysteine residue and varies from +8 % to +1.5 % (see graph 5 of figure 4)
- the cysteine residue is very rare in position 1 and 2 for proteins which have no bridge (see 1 of fig. 4 and 5).

Amino acids that are typical of an environment are listed below :

Amino Acid	*First set*	*Second set*
D	no influence	
L, C	T	F, I_a, I_e
A, D	T, F	I_a, I_e
H, S	I_a	I_e
E, Q, K	I_a	T, F, I_e
T, M	I_e	T, F, I_a

Table 2. Discriminant amino acids

4. Conclusion and future work

The Disulfide Bridge DataBase appears to us as a convenient database for anyone involved in the prediction of disulfide bridge prediction. We think that the growing number of proteins into the DBDB should lead to more precise results. Further analysis have to be performed for example by using information about the secondary structure. Our next step toward disulfide bridge prediction will be to use physico-chemical properties of amino acids in order to characterize the environment of cysteine residues in terms of volume, hydrophobicity, surface area, charge and polarity.

References

[1] Martelli Pier Luigi, Fariselli Piero, Malaguti Luca and Casadio Rita, *Prediction of the disulfide-bonding state of cysteines in proteins at 88 % accuracy*, Protein Science, 11:2735-2739, 2002.

[2] Matsumura M., Signor G. and Mathews B.W., *Substantial increase in protein stability by multiple disulfide-bonds*, Nature, 342, 291-342, 1989.

[3] Berman H.M., Westbrook J., Feng Z., Gilliland,G., Bhat T.N., Weissig H., Shindyalov I.N. and Bourne P.E., *The Protein Data Bank*. Nucleic Acids Res., 28, 235-242, 2000.

[4] Neves Petersen Maria Teresa, Johnson Per Harald and Petersen Steffen B., *Amino acid neighbours and detailed conformational analysis of cysteines in proteins*, Protein Engineering, vol 12, no. 7, pp 535-548, 1999.
[5] Hyunsoo Kim and Haesun Park, *Protein secondary structure prediction based on an improved support vector machines approach*, Protein Engineering vol. 16 no. 8 pp. 553-560, 2003.
[6] Vullo A. and Frasconi P., *Disulfide connectivity prediction using recursive neural networks and evolutionary information*. Bioinformatics,20, 653-659, 2004.
[7] Vinayagam A., Pugalenthi G., Rajesh R. and Sowdhamini R. *DSDBASE: a consortium of native and modelled disulphide bonds in proteins*. Nucleic Acids Res., 32, D200-D202, 2004.
[8] Srinivasan K.N., Gopalakrishnakone P., Tan P.T., Chew K.C., Cheng B., Kini R.M., Koh J.L., *SCORPION, a molecular database of scorpion toxins*. Toxicon, 40, 23-31, 2002.
[9] O'Connor D, Yeates TO., *GDAP: a web tool for genome-wide protein disulfide bond prediction*., Nucleic Acids Res., 32, 2004.
[10] Fariselli P. and Casadio R., *Prediction of disulfide connectivity in proteins*. Bioinformatics, 17, 957-964, 2001.

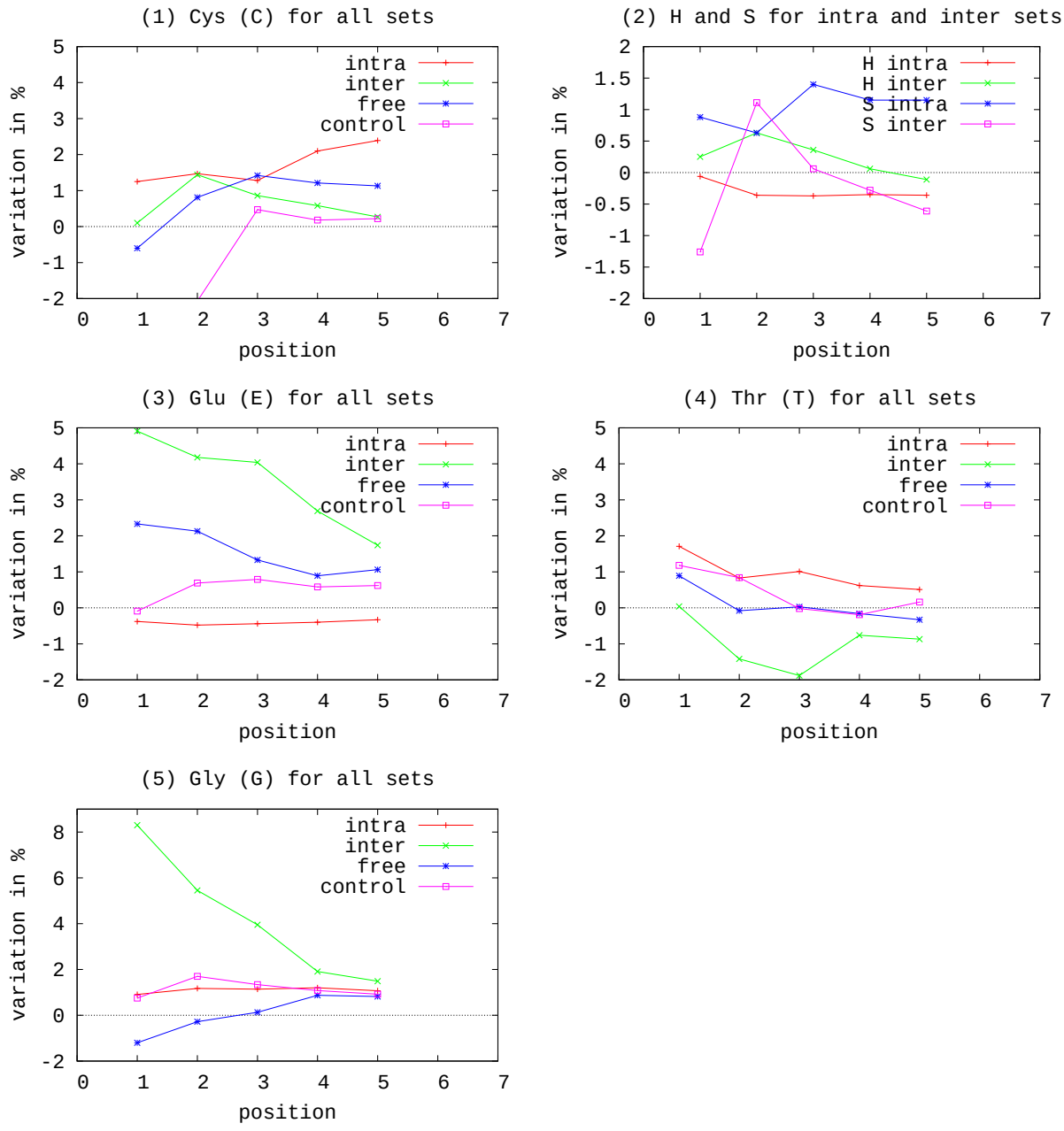
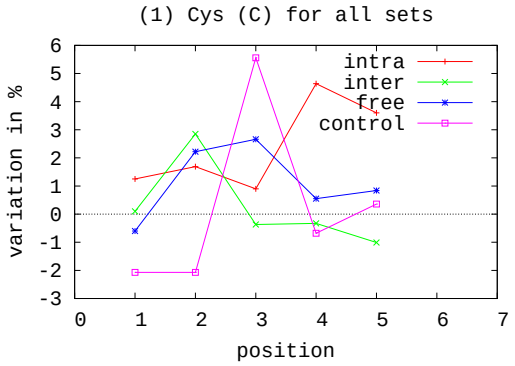
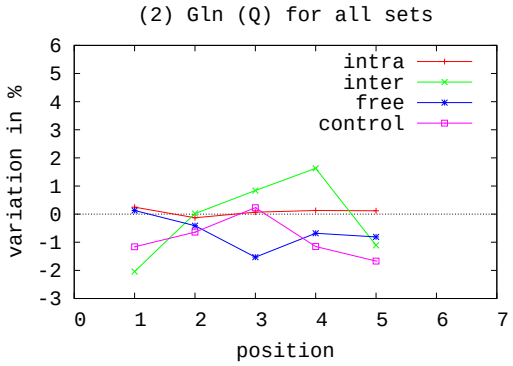


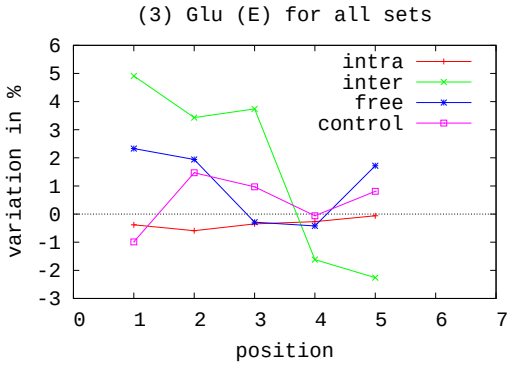
Figure 4. Some variations of frequencies from the cumulative analysis



(a) Cysteine



(b) His and Glu



(c) Gln

Figure 5. Some variations of frequencies from the position-specific analysis

General Information

Id :	153
Name :	Bromelain inhibitor vi
Classification :	Cysteine protease inhibitor
Taxonomy :	Ananas comosus
EC id :	
PDB Identifier(s) :	2BI6
# Chains :	2
# Intra :	2
# Inter :	3
# Intra + Inter :	5
Bibliography :	Hatano, K., Kojima, M., Tanokura, M., Takahashi, K.: Solution Structure of Bromelain Inhibitor Vi from Pineapple Stem To be Published
Remark :	NMR : 18 Structures

Chains

C free Cystein C Intra bond C Inter bond

Chain L

```
         1         2         3         4         5
12345678901234567890123456789012345678901234567890
---------+---------+---------+---------+---------+

TACSECVCPLR
```

Chain H

```
         1         2         3         4         5
12345678901234567890123456789012345678901234567890
---------+---------+---------+---------+---------+

EEYKCYCTDTYSDCPGFCKTCKAEFGKYICLDLISPNDCVK
```

Get Chains in Fasta Format

Bonds

Intra bond 1	H 14 <-> 21 H
Intra bond 2	H 18 <-> 30 H
Inter bond 1	L 3 <-> 7 H
Inter bond 2	L 6 <-> 39 H
Inter bond 3	L 8 <-> 5 H

Figure 6. Sample information for protein 2BI6 which contains 2 chains L and H, 2 intra and 3 inter bonds