

# Prediction of the Number of Residue Contacts in Proteins

Piero Fariselli and Rita Casadio

CIRB Biocomputing Unit  
and

Lab. of Biophysics, Dept. of Biology University of Bologna  
via Imerio 42, 40126 Bologna Italy

Phone: +39 051 2094005 Fax: +39 051 242576

e-mail: [Casadio@alma.unibo.it](mailto:Casadio@alma.unibo.it), [Piero@biocomp.unibo.it](mailto:Piero@biocomp.unibo.it)

## Abstract

Knowing the number of residue contacts in a protein is crucial for deriving constraints useful in modeling protein folding and/or scoring remote homology search. Here we focus on the prediction of residue contacts and show that this figure can be predicted with a neural network based method. The accuracy of the prediction is 12 percentage points higher than that of a simple statistical method. The neural network is used to discriminate between two different states of residue contacts, characterized by a contact number higher or lower than the average value of the residue distribution. When evolutionary information is taken into account, our method correctly predicts 69% of the residue states in the data base and it adds to the prediction of residue solvent accessibility. The predictor is available at <http://www.biocomp.unibo.it>

**Keywords:** Protein structure predictions; protein contacts; neural networks; solvent accessibility; evolutionary information.

## Introduction

A major challenge in molecular biology is the elucidation of the functional properties of proteins in terms of structural and dynamical features. In this context the core problem is posed by the process of protein folding during which the protein settles into a stable and well definite three-dimensional structure. Its knowledge is valuable in determining the structure to function relationship. Moreover it justifies the considerable effort being expended to bridge the gap between the amount of 3D structures known with atomic resolution and the overwhelming quantity of amino acid sequence data (Sánchez and Sali, 1998).

### Solvent accessibility and number of contacts

In the attempt of predicting aspects of protein structural organization, a basic and informative distinction to make is

the degree to which residues in the structure interact with the solvent molecules. The relative solvent accessibility is the one-dimensional descriptor which captures this distinction. Prediction of residue accessibility has been attempted with different methods based on neural networks without (Holbrook et al., 1990) or with evolutionary information (Rost and Sander 1994a), on Bayesian analysis (Thompson and Goldstein 1996) and residue substitution matrices (Pascarella et al., 1998). Recently, a simple statistical approach was introduced, that classifies residues into a buried or non-buried category with no reference to the surrounding sequence (Richardson and Barlow 1999). This "bottom line" method predicts the level of solvent accessibility of a residue, based on the identity of residues, independently of their context. Despite its simplicity (no context dependence), the "bottom line" accuracy is as good as that of the more sophisticated methods, using neural networks and bayesian statistics. It also indicates that the residue propensity to be exposed or not exposed to the solvent is sufficient for the prediction score. A comparison of all the available methods shows that the accuracy value levels around 69-71%, when single protein sequence is used (Richardson and Barlow 1999).

Predictors of solvent accessibility predict relative accessibility classes. This is done using the computed solvent accessibility from DSSP program (Kabsch and Sander 1983) and normalizing it at the maximum value of exposed surface area obtainable for each residue. Different arbitrary threshold values of solvent accessibility are chosen to define binary categories (buried and exposed) or ternary categories (buried, partially exposed, or exposed).

Likewise secondary structure prediction (Rost and Sander, 1994b), approaches to predicting solvent accessibility benefit from using evolutionary information. An increase of 5 percentage points in the prediction accuracy was reported both with neural networks (Rost and Sander 1994a) and Bayesian methods (Thompson and Goldstein 1996).

Another one-dimensional descriptor basic to protein structural information is the number of stabilizing contacts that residues make in the protein folded globule (for review see Dill 1999). Based on the notion that less exposed residues are preferentially involved in hydrophobically

driven chain compaction, solvent accessibility has been routinely used to evaluate also the number of residue contacts. In order to simulate the hydrophobic collapse in model proteins the number of residue contacts is chosen as the inverse measure of the residue solvent accessibility and in the case of simple lattice protein models, it is the only source of interaction (Sali et al., 1994).

In this paper we show that although a strict connection between accessibility and contact number is commonly accepted, residue surface accessibility is differently distributed than the number of residue contacts in a data base of selected proteins and that residue classification may be different depending on which property is highlighted. Therefore a direct prediction of the number of residue contacts is worth in many cases.

### Relevance of residue contact information

Knowing the correct positions of residue contacts in proteins has been proven extremely useful to determine the three-dimensional structure of a given protein, as it was recently demonstrated in the CASP3 competition (CASP3, Ortiz et al., 1999). Moreover, when remote homology is searched, it is very profitable to derive a surface potential based on the distribution of contact numbers for each residue. This is computed by implementing the inverse of the Boltzman rule (Flöckner et al., 1995) or by using the notion of contacts among residues to improve existing threading algorithms (Olmea et al., 1999).

In an off-lattice context the number of contacts for each residue is computed inside a spherical cut-off centered into each residue and by counting the number of residues falling inside a defined volume (Flöckner et al., 1995).

In the last few years different attempts to predict contacts (Shindyalov et al., 1994; Olmea and Valencia, 1997; Fariselli and Casadio 1999) and distances among residues in proteins (Aszodi et al., 1995; Lund et al. 1997; Gorodkin et al., 1999) have been made with some extent of success.

In this paper we develop a predictor capable of discriminating if a given residue, depending on its sequence context, has a number of contacts greater or lower than its average value in the data base. This type of classification is complementary to predicting residue solvent accessibility and can be used to improve methods suited to predict protein structure.

## The Method

### The protein data base

Neural networks are trained and tested on a database of proteins selected from the Protein Data Bank using the PDB\_select algorithm (Hobohm et al., 1992). For the training phase proteins with an identity value <25% were extracted from the PDB\_select\_oct\_1997 file

(<http://www.embl-heidelberg.de>). This set was then reduced by excluding those chains whose backbone is interrupted. The final set is listed in Appendix 1 using the PDB acronyms.

When single sequences were used, inputs of the predictor were derived from the PDB files; in the case of multiple sequence inputs, the sequence profiles generated with the MaxHom program were extracted from the HSSP files (Schneider and Sander, 1991).

### Computation of residue solvent accessibility and contact number

Residue solvent accessibility is evaluated by using the DSSP program (Kabsch and Sander 1983). The value is normalized to the maximal exposed surface area of each residue (Rose et al., 1985) in the data base of selected proteins (Appendix 1).

The number of inter-residue contacts for each residue of the data base is computed by defining a spherical protein volume centered in the  $C\beta$  atom (or  $C\alpha$  for GLY) and with a radius equal to 6.5 Å.

### The predictor and the measure of accuracy

Initially, for each residue the frequency distribution of the number of contacts is computed using the protein data base. A standard feed-forward neural network based predictor is then trained (and tested) to classify whether a given residue, depending on the context of the input window, has a number of contacts lower or higher than its average distribution value. The training procedure is performed using single sequence or sequence profile as input to the networks.

The training algorithm is back-propagation (Rumelhart et al. 1986). The network architecture consists of a perceptron with one hidden layer, and two output neurons. The number of hidden neurons was changed from 2 to 32 without significantly affecting the predictive performance. The input window was 1 to 15 residue long, depending on the test case. The predictor implemented with a 1 residue long window was used as a "bottom line" reference. This simple predictor always assigns a residue to its most abundant class independently of its environment (Richardson and Barlow, 1999).

A cross validation procedure was adopted by splitting the proteins listed in Appendix 1 into 10 subsets of almost equal size.

In order to score the efficiency we used the following accuracy indices (Fariselli et al. 1993): Q2 is the number of correctly predicted residues divided by the total number of residues; PC is the number of correct assignments to a given class divided by the number of all the residues predicted in that class; Q is the number of correct assignments to a given class divided by the total number of observed in that class; C is the Matthews' correlation coefficient.

## Results and Discussion

### Residue solvent accessibility and distribution of contact numbers

A key point of our work is to elucidate the difference between solvent accessibility and number of contacts of a given residue. To this purpose, we first computed the distributions of solvent accessibility and of contact numbers for each residue in the protein data base (Figure 1 and 2, respectively). The solvent accessibility distributions are characterized by three patterns: the first, typical of the majority of residues (hydrophobic and polar) is characterized by higher frequencies of occurrence corresponding to low values of solvent accessibility (in the range of 10%); the second, comprising charged residues (except lysine (K)) and glutamine (Q) shows higher frequency values picking both around 10% and around 40-60% accessibility values, respectively; the third typical of lysine (K) is characterized by frequencies of occurrence peaking around 40%. These data, which are in agreement with other authors' statistics (Holbrook et al., 1990; Richardson and Barlow 1999), suggest that a volumetric effect is dominant over the single residue chemico-physical characteristics (in a protein, on average, more residues are buried than exposed on the surface, Janin 1979).

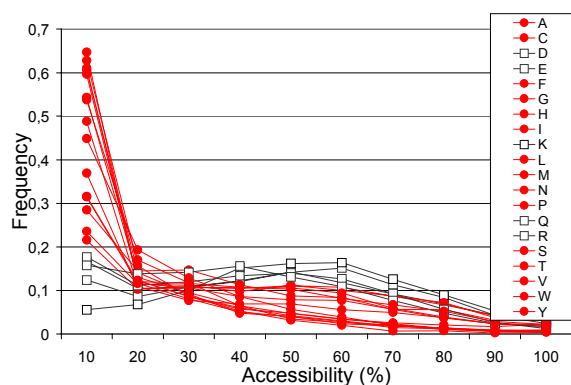


Figure 1. Frequency distribution of the relative solvent accessibility for the 20 residues. The residues are depicted using two different symbols depending on the shape of their distribution: filled circles when the distribution has only one pick in the low accessibility region and open squares when it is bimodal or flat.

Alternatively, the distributions of residue contacts are characterized by a unique pattern (roughly bell-shaped) independently of the residue type (Figure 2). Distributions can be apparently distinguished depending on the two different positions of their maximal frequency value (equal to the distribution average value). The distributions of hydrophobic residues, with the only exception of tyrosine (Y), peak at the highest average value of contacts (6-7) whereas the distributions of polar and charged residues peak at the lowest average value (4-5). Given the

distribution shape, one can speculate that each residue (or its  $C\beta$  atom) has a preferred coordination (its average value of contacts) around which it fluctuates due to the protein environment. For this reason a mean force potential based on the residue contacts can help in the remote homology search (Flökner et al., 1995).

For each protein of the data base a correlation is computed between the number of residue contacts and the residue solvent accessibility of the chain. The correlation coefficient values obtained, considering both different contact volumes and different cut-off values of solvent accessibility, indicate that correlation between the two descriptors is poor (Table 1). We can conclude that predicting the number of contacts of a given residue is different from predicting its solvent accessibility.

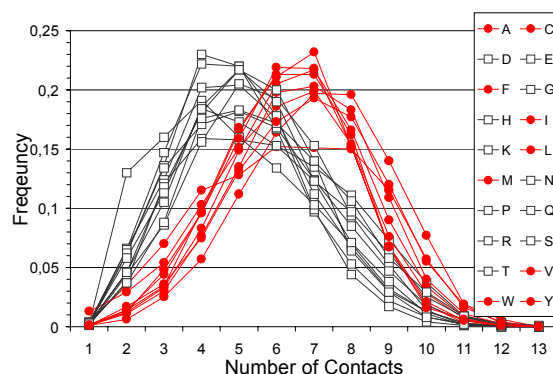


Figure 2. Frequency distribution of the number of contacts for the 20 residues. The residues are plotted using two different symbols depending on whether the mean value of their contacts is peaking around 4 (open squares) and around 7 (filled circles).

Table 1. Correlation coefficients between the numbers of residue contacts and the relative accessibility of the proteins in the data base.

Accessibility Cut-off	9%	16%	50%
$C\beta$	0.42	0.44	0.39
$C\alpha$	0.35	0.37	0.33

Contact numbers are computed using  $C\beta$  or  $C\alpha$  atoms as volume centers and a radius of 6.5 Å. Binary categories (buried or exposed) were discriminated by choosing three different percentage threshold values of solvent accessibility.

### The predictor at work

The best performing architecture of the predictor was selected both by changing the number of hidden neurons (from 2 to 32) and the window input dimension (odd numbers from 1 to 15). The accuracy of the predictor, at a fixed number of hidden neurons (4), is shown in Figure 3 as a function of the input window length. A relevant increase on the accuracy value is noticeable when the

window length is enlarged from 1 to 5 residues. The accuracy increases by about 5 percentage points in passing from 1 to 3 residues, and a further 2% is gained with a 5 residue long window. However when larger windows ranging from 7 to 15 residues are used, the accuracy is slightly affected. In this range, as already pointed out for the solvent accessibility prediction, the choice of the window length has a marginal influence on the results (Rost and Sander 1994a). The same observation holds when the number of hidden neurons at optimal window length is increased: the efficiency is not significantly affected by increasing the number of hidden neurons above 4.

It is however noticeable (Figure 3) that the use of multiple sequence alignment in the form of profile sequence (MS) as input to the network increases the predictor accuracy of three percentage points as compared to single sequence (SS). This indicates that the number of contacts in proteins is to a certain extent a conserved property. Indeed it is evident that the accuracy is much lower when the window length of the predictor is one and the prediction is totally context independent.

The base line predictor and the best performing predictors, either using single sequence or multiple

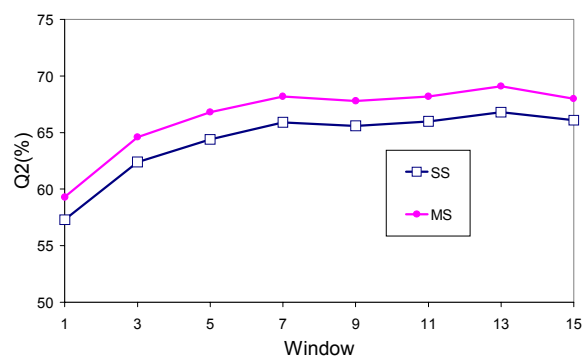


Figure 3. Testing accuracy of the neural network based predictor as a function of the length of the input window. The networks were trained with single sequence (SS) or multiple sequence (MS) as input and using 4 hidden neurons.

sequence, are compared by evaluating different scoring indexes (Table 2). The two categories discriminated are labeled H (higher than the average) and L (lower or equal to the average). It is evident that both the single sequence (SS) and multiple sequence (MS) predictors score higher (up to 12 percentage points with evolutionary information) than the baseline predictor (BASE).

To our knowledge this is the first attempt to predict the number of inter-residue contacts in proteins. The use of neural networks was prompted by the indication that for a related task (prediction of solvent accessibility) Bayesian and neural network-based methods perform similarly. We find that the number of inter-residue contacts in proteins can be predicted with a good accuracy provided that context and evolutionary information are taken into account.

Table 2. Scoring of the neural network based predictor

Method	Q2	Q(L)	Q(H)	PC(L)	PC(H)	C
BASE	0.57	1.00	0.00	0.57	0.00	0.00
SS	0.66	0.78	0.48	0.68	0.61	0.28
MS	0.69	0.81	0.51	0.70	0.66	0.35

BASE=baseline method; SS=neural network with single sequence as input; MS=neural network with multiple sequence as input; Q2= overall accuracy; Q(L) and Q(H) = accuracy normalised to the observed residue for the classes L (lower and equal than the average) and H (higher than the average) respectively; PC(L) and PC(H) = accuracy normalised to the predicted residue for the L and H classes, respectively; C = correlation coefficient. Indexes are computed adopting a cross validation procedure.

## Acknowledgements

Financial support to this work was provided by a grant of the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) delivered to the project “Structural, Functional and Applicative Prospects of Proteins from Thermophiles” and by a grant for a target project in Biotechnology of the Italian Centro Nazionale delle Ricerche (CNR)

## Appendix 1. The protein data base

1191_	1531_	1a0aa	1a0b_	1aa0_	1aa2_	1aa3_
1aaf_	1aaya	1ab3_	1ab8a	1aba_	1abrb	1acp_
1ad2_	1ads_	1aerb	1afp_	1afra	1afwa	1ag2_
1ag4_	1agna	1agqd	1agre	1ah7_	1ah9_	1aho_
1ai6a	1aie_	1aiha	1aikc	1aj3_	1ajj_	1ajya
1ak0_	1ak1_	1akjd	1ako_	1akz_	1alo_	1aly_
1amm_	1amp_	1an9a	1anu_	1ao7b	1aoca	1aoqa
1aoy_	1aoza	1ap8_	1apf_	1apj_	1apyb	1aq0a
1aq6a	1aqb_	1aqt_	1arla	1arb_	1ark_	1arv_
1as4b	1ash_	1asx_	1asya	1atib	1atla	1atzb
1avma	1avob	1awcb	1awd_	1awj_	1axib	1axn_
1bak_	1bbpa	1bcmb	1bcn_	1bct_	1bdma	1bdo_
1beba	1benb	1beo_	1bfg_	1bfma	1bfta	1bgk_
1bgp_	1bhmb	1bip_	1bkf_	1ble_	1bmfq	1bnb_
1bncb	1bnda	1bor_	1bova	1bp1_	1brn1	1broa
1btb_	1btma	1btn_	1bvh_	1bvpl	1bw3_	1byb_
1c5a_	1cby_	1cdb_	1cdi_	1cds_	1cem_	1cewi
1cex_	1cfb_	1cfe_	1cfh_	1cfya	1chd_	1chka
1chma	1cid_	1clc_	1cmke	1cmr_	1cmyb	1cne_
1cnt2	1cnv_	1crka	1csbb	1csga	1csh_	1csn_
1ctj_	1cto_	1cur_	1cyda	1cyx_	1d66a	1daaa
1dad_	1ddf_	1deaa	1dec_	1def_	1delb	1dhr_
1div_	1djxa	1dkgb	1dka_	1dkza	1dlha	1doka
1dora	1dpga	1dru_	1dubb	1dupa	1dxy_	1eal_
1ebpa	1eca_	1ecea	1ecmb	1ecpa	1ecra	1ede_
1edg_	1edmb	1edt_	1ehs_	1erd_	1erv_	1esc_
1esfa	1etpa	1eur_	1exh_	1ezm	1fbr_	1fc1a
1fcda	1fdm_	1fjma	1flel	1fmbt	1fna_	1foka
1fpka	1ftpa	1fua_	1furb	1fvka	1fw_	1gai_

1garb 1gcb\_ 1gdob 1ggga 1gifa 1gin\_ 1gky\_  
 1glef 1gnd\_ 1gnha 1gnwa 1goh\_ 1gpb\_ 1gpc\_  
 1gpm 1gpt\_ 1gsa\_ 1gtma 1gtqa 1gtra 1guqb  
 1gvp\_ 1gyla 1hava 1hcd\_ 1hcg 1hcr 1hev\_  
 1hfc\_ 1hgb 1hja 1hla 1hloa 1hoe\_ 1hqi\_  
 1hrya 1hsba 1hsn\_ 1hsta 1htmb 1htp\_ 1htp  
 1hula 1hwah 1idaa 1idk\_ 1ido\_ 1if1b 1ifc\_  
 1ife\_ 1igd\_ 1igna 1ihfa 1inp\_ 1ipsa 1ipwb  
 1irk\_ 1irl\_ 1irsa 1isua 1itbb 1iva\_ 1ixh\_  
 1iyv\_ 1jaca 1jdw\_ 1jer\_ 1jeta 1jhga 1jkw\_  
 1jli\_ 1jmca 1jpc\_ 1jrhi 1jsuc 1jvr\_ 1jxpa  
 1kaz\_ 1kid\_ 2kinb 1kit\_ 1knb\_ 1knya 1kpf\_  
 1krt\_ 1ksr\_ 1kte\_ 1kuh\_ 1kul\_ 1kvt\_ 1kzub  
 1lam\_ 1latb 1lba\_ 1lbu\_ 1lcl\_ 1lct\_ 1leb\_  
 1lfb\_ 1lgha 1lis\_ 1lki\_ 1lkka 1lmb4 1lpba  
 1lfn\_ 1lrv\_ 1lt5d 1ltsa 1lucb 1lxa\_ 1lxta  
 1mai\_ 1mak\_ 1mbd\_ 1mhlc 1mkaa 1mla\_ 1mml1  
 1mnm 1mola 1mpga 1mrj\_ 1mrc\_ 1msec  
 1mspb 1muca 1mupa 1mzm\_ 1nbab 1nbba 1nbca  
 1ncib 1ngr\_ 1nif\_ 1nkl\_ 1nls\_ 1noe\_ 1nox\_  
 1noya 1npk\_ 1npoc 1nre\_ 1nsgb 1nsya 1nula  
 1nwa 1nxb\_ 1obpa 1ocp\_ 1onra 1opd\_ 1opr\_  
 1ospo 1otfa 1otga 1oyc\_ 1p38\_ 1pce\_ 1pcf  
 1pdc\_ 1pdgc 1pdnc 1pdo\_ 1pea\_ 1pex\_ 1pf  
 1pft\_ 1pgs\_ 1phc\_ 1php\_ 1pih\_ 1pii\_ 1pioa  
 1ppk\_ 1plc\_ 1plr\_ 1pmi\_ 1pne\_ 1pnkb 1poa  
 1poc\_ 1pot\_ 1pou\_ 1ppn\_ 1ppt\_ 1prcc 1prea  
 1prr\_ 1psla 1ptq\_ 1pud\_ 1put\_ 1pyab 1pyda  
 1pyp\_ 1pysa 1qapa 1qnf\_ 1quf\_ 1r69\_ 1ra9\_  
 1rcf\_ 1regy 1res\_ 1rgea 1rgs\_ 1rie\_ 1rlw\_  
 1rmd\_ 1rof\_ 1roo\_ 1rpo\_ 1rro\_ 1rsy\_ 1rtna  
 1rusa 1rvaa 1ryt\_ 1sbp\_ 1sera 1sfe\_ 1sfta  
 1sgpi 1shca 1skye 1skz\_ 1slta 1slua 1slu  
 1smea 1smna 1smpi 1smtb 1smvc 1sqc\_ 1sra\_  
 1sro\_ 1svb\_ 1svpa 1svq\_ 1tabi 1tada 1taha  
 1tc3c 1tca\_ 1tdtc 1tfb\_ 1tfe\_ 1tfpa 1thja  
 1thtb 1thv\_ 1thx\_ 1tib\_ 1tif\_ 1tih\_ 1tiid  
 1tiv\_ 1tle\_ 1tlk\_ 1tml\_ 1tnra 1tpm\_ 1trka  
 1tsq\_ 1tul\_ 1tum\_ 1tupc 1tvxa 1uae\_ 1ubi\_  
 1uby\_ 1udii 1ulp\_ 1unka 1urna 1utg\_ 1uxd\_  
 1luxy 1vba4 1vcaa 1vcc\_ 1vdfa 1vhh\_ 1vhp\_  
 1vhra 1vif\_ 1vig\_ 1vin\_ 1vls\_ 1vmoa 1vnc  
 1vpsa 1vsd\_ 1vtx\_ 1vvc\_ 1vwld 1wba\_ 1wdca  
 2wea\_ 1wer\_ 1whi\_ 1who\_ 1wiu\_ 1wsyb 1wtua  
 1xbrb 1xdtr 1xgsa 1xikb 1xnb\_ 1xxca 1xyza  
 1yaia 1yasa 1ycc\_ 1ycqa 1ycsb 1ysc\_ 1ysth  
 1ytba 1ytfc 1ytw\_ 1yub\_ 1zaq\_ 1zdd\_ 1zid\_  
 1zin\_ 1znba 1zwd\_ 1zxq\_ 256ba 2abd\_ 2abk\_  
 2acy\_ 2arcb 2ayh\_ 2baa\_ 2bbva 2bds\_ 2bi6h  
 2bopa 2cba\_ 2ccya 2cdx\_ 2chsa 2ctc\_ 2cyp\_  
 2dkb\_ 2dri\_ 2drpa 2dyna 2ech\_ 2end\_ 2erl\_  
 2ezh\_ 2ezk\_ 2fha\_ 2fiva 2fow\_ 2fsp\_ 2gdm\_  
 2hbg\_ 2hp8\_ 2hpda 2i1b\_ 2i16\_ 2ilk\_ 2ktx\_  
 2lgsa 2liv\_ 2masa 2mcm 2mpra 2msbb 2mtac  
 2naca 2ncm\_ 2nef\_ 2omf 2pac\_ 2pgd\_ 2phla  
 2phy\_ 2pii\_ 2plda 2pola 2por\_ 2ptd\_ 2pth\_  
 2rn2\_ 2rslc 2rspb 2sak\_ 2scpa 2sici 2sil\_  
 2sn3\_ 2sns\_ 2spca 2stv\_ 2stwa 2tbd\_ 2tgi\_  
 2tysa 2ucz\_ 2vgh\_ 2vhba 2vik\_ 3b5c\_ 3chy\_  
 3cla\_ 3cyr\_ 3grs\_ 3lzt\_ 3mdda 3pbga 3pchm  
 3pte\_ 3r1ra 3sdha 3ulla 4aaha 4dpvz 4hmga  
 4htci 4mt2\_ 4pgaa 4pgmb 4rhv1 4xis\_ 4csma  
 5hpga 5icb\_ 5nul\_ 5p21 5pti\_ 5znf\_ 5gsva  
 7ahla 7gata 7rsa\_ 8abp\_ 8atcb 8ruc1 8rxna

## References

Aszodi, A., Gradwell, M.J., and Taylor, W.R. 1995. Global fold determination from a small number of distance restraints *J. Mol. Biol.* 251:308-326.

CASP3 <http://predictioncenter.llnl.gov/casp3/Casp3.html>

Dill, K.A. 1999. Polymer principles and protein folding. *Protein Sci.* 8:1166-1180.

Fariselli, P., Compiani, M., Casadio, R. 1993. Predicting secondary structures of membrane proteins with neural networks *Eur Biophys J* 22: 41- 51.

Fariselli, P., and Casadio, R. 1999. Neural network based predictor of residue contacts in proteins. *Protein Engng* 12:15-21

Flökner, H., Braxenthaler, M., Lackner, P., Jaitz, M., Ortner, M., and Sippl M. J. 1995. Progress in fold recognition. *Proteins* 3:376-386.

Gorodkin, J., Lund, O., Andersen, C.A., and Brunak, S. 1999. Using Sequence Motifs for Enhanced Neural Network Prediction of Protein Distance Constraints. In: *Proc. of the seventh international conference on Intelligent Systems for Molecular Biology (ISMB '99)*, 95-105 AAAI Press.

Hobohom, U., Scharf, M., Schneider, P. and Sander, C. 1992. Selection of representative protein data sets. *Protein Sci.* 1: 409-417.

Holbrook, S. R., Muskal, S. M. and Kim S. H. 1990. Predicting surface exposure of amino acids from protein sequence. *Protein Engng* 3:659-665.

Janin, J. 1979. Surface and inside volumes in globular proteins. *Nature* 277:491-492.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577-2637.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., and Brunak S. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Protein Engng* 10:1241-1248.

Olmea, O., and Valencia, A. 1997. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.* 2:S25-32.

Olmea, O., Rost, B., Valencia, A. 1999. Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* 293:1221-1239.

Ortiz, A.R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., and Skolnick J. 1999. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins* Suppl 3:177-185.

Pascarella, S., De Persio, R., Bossa, F., and Argos, P. 1998. Easy method to predict solvent accessibility from multiple protein sequence alignments. *Proteins* 32:190-199.

Richardson, C.J., and Barlow, D. J. 1999. The bottom line for prediction of residue solvent accessibility. *Protein Engng* 12:1051-1054.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* 229:234-238.

Rost B. and Sander C. 1994a. Conservation and prediction of solvent accessibility in protein families. *Proteins* 20:216-226.

Rost B. and Sander C. 1994b. Combining evolutionary information and neural networks to predict secondary structure of proteins. *Proteins* 19:55-72.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. 1986. Learning representation by back-propagation error. *Nature* 323: 533-537.

Sánchez, and Sali, 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *P.N.A.S.* 954:13597-602

Sali, A., Shakhnovich, E., and Karplus, M. 1994. How does a protein fold? *Nature* 369:248-251

Schneider, R., Sander, C. 1991. Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 9: 56-68

Shindyalov, I.N., Kolchanov, N.A., and Sander, C. 1994. Can three-dimensional contacts of proteins be predicted by analysis of correlated mutations? *Protein Engng* 7:349-358.