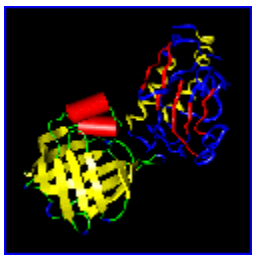


ANOLEA: ATOMIC NON-LOCAL ENVIRONMENT ASSESSMENT

Francisco Melo and Ernest Feytmans.

Laboratory of Structural Molecular Biology, University of Namur, Belgium.

Assessing Protein Structures

The assessment of protein structures is a delicated matter. Currently, there is not a single method able to consistently and accurately predict the three-dimensional structure of a protein. Moreover, our lack of knowledge about protein folding and stability makes difficult to develop such a method.
Similarly, there is no single method able to consistently and accurately predict the errors in a protein structure. For that reason, we strongly recommend to use as many as possible different methods when assessing a protein structure (below we provide links to some well- known and widely used methods). Different methods use different approaches, thus they can complement to each other. In that way, you can have more confidence about the predicted error of specific regions in the protein.

A List of 'Well-known' and 'Widely-used' Methods to Assess Protein Structures

- PROCHECK
- PROSA II
- SURVOL
- VERIFY3D
- WHAT CHECK

What is ANOLEA ?

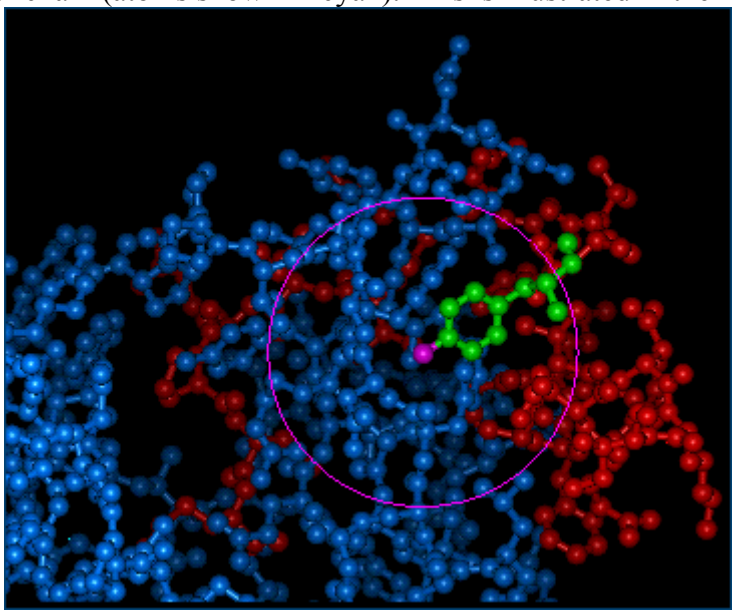
ANOLEA (**A**tomic **NO**n-**L**ocal **E**nvironment **A**ssessment) is a server to assess protein structures. It performs energy calculations on a protein chain, evaluating the "non-local environment" (NLE) of each heavy atom in the molecule. The NLE of one heavy atom is defined as all the heavy atoms, within an Euclidean distance of 7 Å , that belong to an amino acid that is farther than 11 residues in the chain or that belong to another chain. The energy of each atom pairwise interaction is taken from a distance-dependent knowledge-based mean force potential (DD-KB-MFP) that has been derived from a

database of 147 non-redundant proteins chains with a sequence identity below 25% and solved by x-ray crystallography with a resolution lower than 3 Å . If you want to see the organization of this server click here.

ANOLEA (**A**tomic **No**n-**L**ocal **E**nvironment **A**ssessment) is a server that performs energy calculations on a protein chain, evaluating the "**N**on- **L**ocal **E**nvironment" (NLE) of each heavy atom in the molecule. The energy of each pairwise interaction in this non-local environment is taken from a distance-dependent knowledge-based mean force potential that has been derived from a database of 147 non-redundant protein chains with a sequence identity below 25% and solved by X-Ray crystallography with a resolution lower than 3 Å .

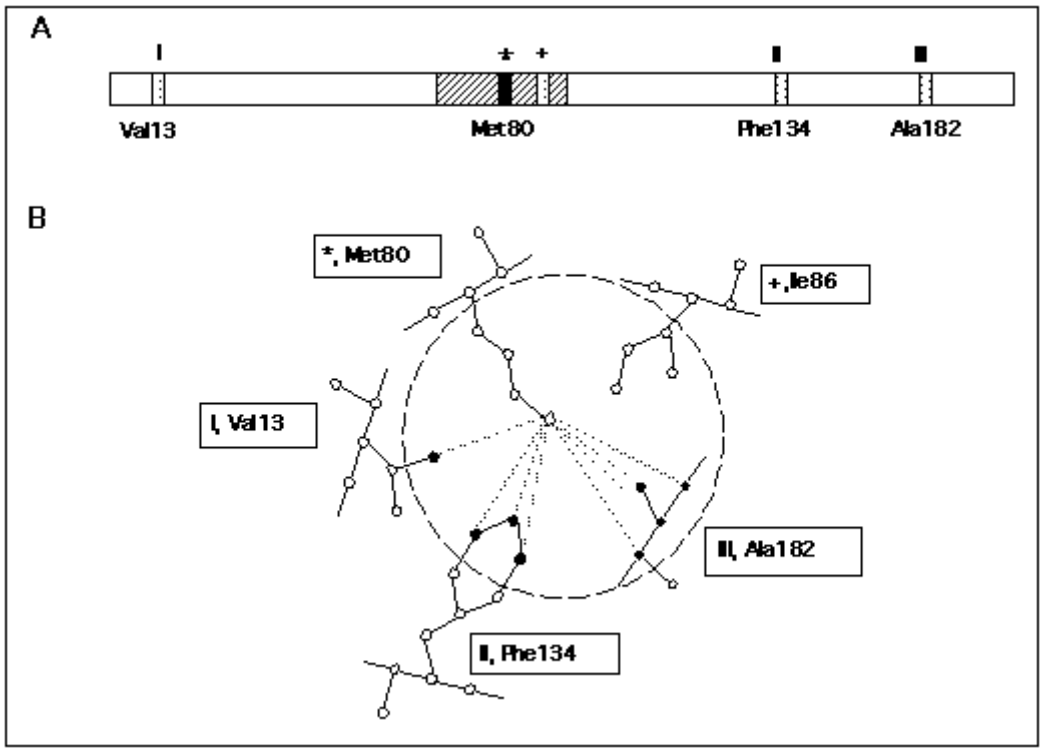
Non local environment

The "non local environment" (**NLE**) of one heavy atom (atom shown in magenta) is defined as all the heavy atoms, within an Euclidean distance of 7 Å , that belong to an amino acid that is farther than 11 residues in the chain (atoms shown in blue) or that belong to another chain (atoms shown in cyan). This is illustrated in the Figure below.



The atoms shown in red and green are not considered in the calculation. These atoms belong to an amino acid that is closer than 12 residues in the chain.

Non local environment scheme

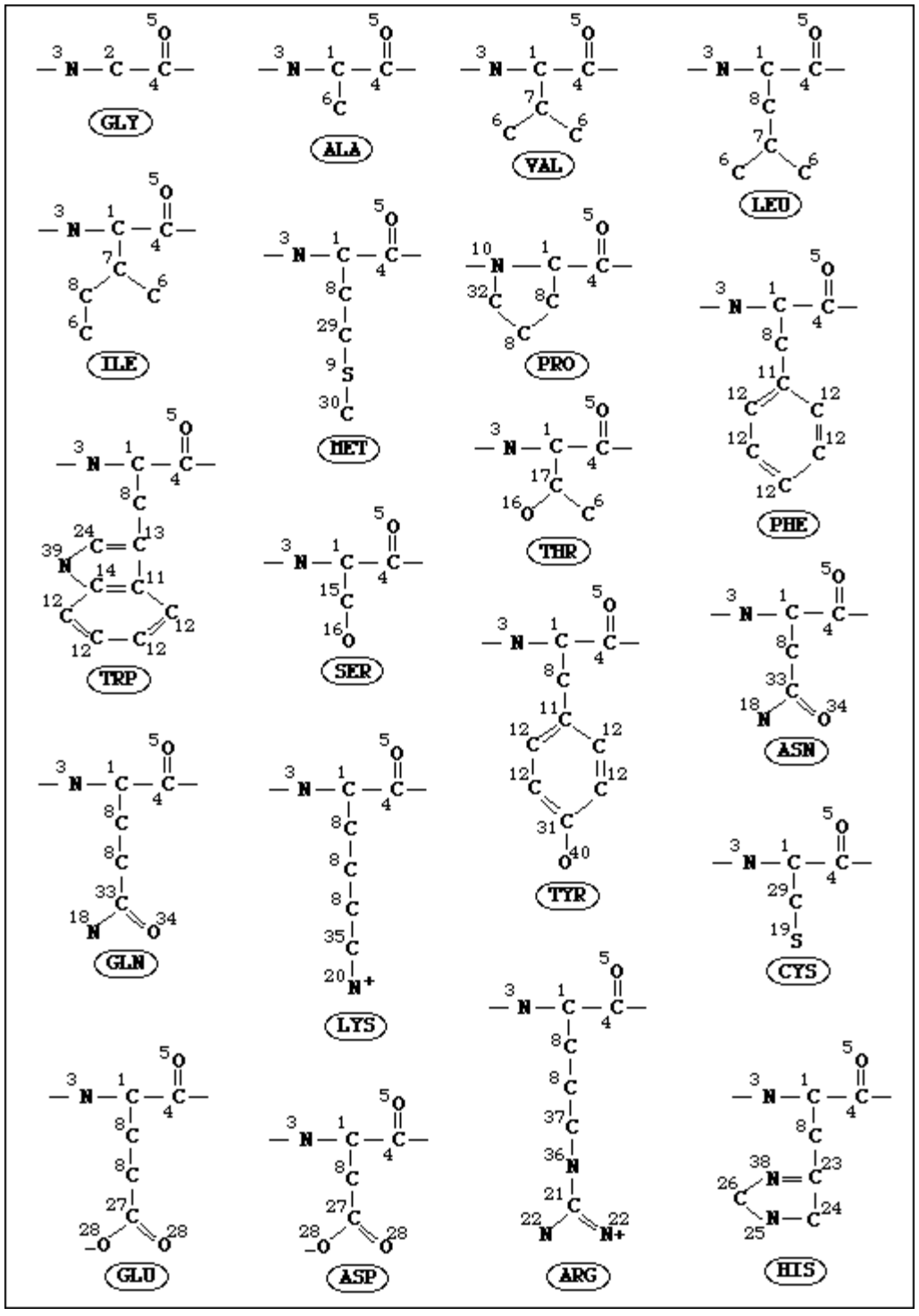


Non-local environment of an atom. (A) Linear representation of a protein sequence. The asterisk represents the amino acid (Met80) containing the atom for which non-local interaction energies are to be evaluated (Ce). The hatched area in the sequence represents the amino acids that are considered as local for Met80 (i.e. not farther than 11 residues in the sequence) and that are not considered in the calculation. Regions I, II and III represent non-local interacting amino acids for Met80, because some of their atoms are closer than 7 Å to Ce of Met80. (B) Two-dimensional structural representation of part of the protein chain illustrated in A. The dashed circle represents a sphere of 7 Å radius centered on Ce of Met80. All the atoms inside this sphere that belong to Met80 (Cb, Cg, Sd) and the local amino acid Ile86 (Cb, Cg1, Cg2, Cd) are shown as open circles and are not considered in the calculation. All the atoms of non-local amino acids that are located within the sphere (Val13, Cg2; Phe134, Ce1, Ce2, Cx; Ala182, N, Ca, C, Cb) are shown as filled circles. All these atoms are included in the non-local interaction energy calculation of the Met80 Ce atom. The Euclidean distances between each one of these atoms and the Met80 Ce atom are calculated, and the corresponding energy values are taken from the energy functions of the atomic mean force potential. The total energy of the Met80 Ce atom is the sum of all the pairwise non-local interaction energy values.

Distance-dependent knowledge-based mean force potential

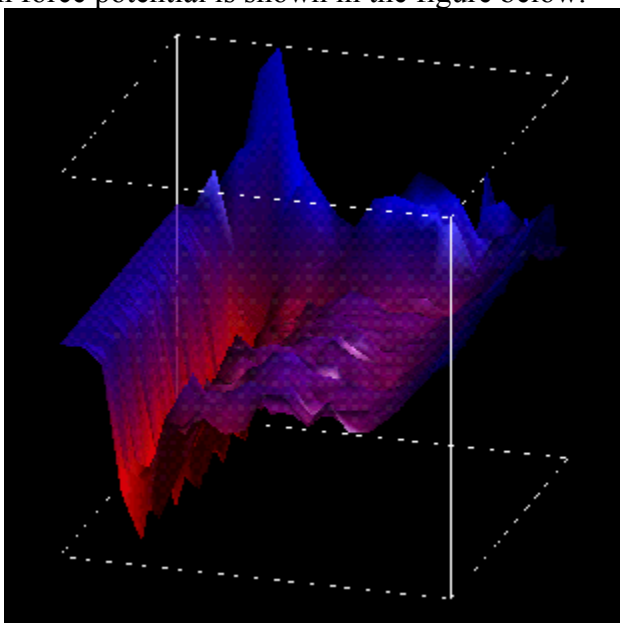
Atom type definitions

ANOLEA uses a mean force potential at the atomic level, which is based on a particular definition of atom types. Forty different atom types are defined for all the heavy atoms of the 20 existant amino acids. In a strict physico-chemical point of view, all the atoms with different environments, connectivities and chemical nature, would be different. In the 20 amino acids the total number of heavy atoms is 167 and the number of non-equivalent heavy atoms is 98. We defined a total number of 40 different atom types for all the heavy atoms of the 20 amino acids (see figure below). The atom type definition is based on its connectivity, chemical nature and location level (side chain or backbone). For example, one type of heavy atom is the carbon of a methyl group bonded to a carbon with sp3 hybridization. This type represents the beta carbon of alanine side chain, the gamma1 and gamma2 carbons of valine, the delta1 and delta2 carbons of leucine, etc.


GLY
ALA
VAL
LEU
ILE
MET
PRO
PHE
TRP
SER
THR
ASN
GLN
LYS
TYR
CYS
GLU
ASP
ARG
HIS

Mean force potential derivation

This mean force potential was calculated to evaluate the non-local environment of each atom in a protein molecule. The non-local environment of any given atom is defined as all the heavy atoms that are included within a 7 Å radius sphere centered on the given atom, and that belong to amino acids that are more distant than 11 residues in the same chain or that belong to another chain. This definition is based on the observation of the energy curves of a previous mean force potential that included local and non-local interactions, where sequence separations higher than 11 do not affect the shape of the energy functions for the different pairwise interactions. As an example, one energy curve of this previous mean force potential is shown in the figure below.



The current potential has been derived using 35 different classes of distance (***l***), ranging from 0 to 7 Å by steps of 0.2 Å . All pairwise ocurrences out of this range were excluded. We considered all the pairs of atoms ocurring within an Euclidean distance of 7 Å , that belong to amino acids that are farther than 11 residues in the chain or that belong to different chains (if you want to see a scheme of this click here). A sequence separation of 2 means that two amino acids are adjacent in the sequence.
The potential was obtained from a database of 147 non-redundant protein chains. The calculation was performed over all these chains. In the case of multimers, all the atom contacts with the other chains were considered.

We have compiled each pairwise contact symmetrically, obtaining 820 (40x41/2) different atom pairwise distributions. The calculation of pairwise pseudo-energy terms has been carried out as described by Sippl (1990). The following expression was used:

$$\Delta E^{ij}(l) = RT\ln\left[Mij \times \sigma\right] - RT\ln\left[Mij \times \sigma \times \frac{f^{ij}(l)}{f^{xx}(l)}\right]$$

where $\mathbf{M_{ij}}$ is the total number of observations for the atomic pair **ij** and corresponds to:

$$Mij = \sum_{l=1}^{n} f(i,j,l)$$

sigma represents the weight given to each observation. We have used **sigma** = 1/50 as proposed by Sippl (1990), such that for 50 observations, $\mathbf{F_{ij\ (l)}}$ and $\mathbf{F_{xx\ (l)}}$ have the same weight in the calculation of $\mathbf{E_{ij\ (l)}}$.

$\mathbf{F_{ij\ (l)}}$ is the relative frequency of ocurrence of the atomic pair **ij** in the class of distance **l** and corresponds to:

$$f^{ij}(l) = \frac{f(i,j,l)}{Mij}$$

$\mathbf{F_{xx(l)}}$ is the relative frequency of ocurrence of all the atomic pairs in the class of distance **l** and can be expressed as:

$$f^{xx}(l) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} f(i,j,l)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} f(i,j,l)}$$

The temperature was set to 293K, so that **RT** is equal to 0.582 kcal/mol.

Bibliography

- Sippl, M.J. (1990) "Calculation of conformational ensembles from potentials of mean force". *J. Mol. Biol.* **213**, 859-883.
- Melo, F. and Feytmans, E. (1997) "Novel knowledge-based mean force potential at atomic level". *J. Mol. Biol.* **267**, 207-222.
- Melo, F. and Feytmans, E. (1998) "Assessment of protein structures based on the non-local energy". *J. Mol. Biol.* **277**, 1141-1152.

Non-redundant dataset

A set of 147 protein chains with complete atomic coordinates was used to perform the calculations. This set was selected from the October 1995 release of PDB representative list (Hobohm et al., 1992; Hobohm & Sander, 1994), excluding all the proteins with duplicated or missing atoms, structural gaps, or with a total number of residues lower than 100. All these proteins share a sequence identity below 25% and have been solved by X-ray crystallography with a resolution better than 3.0 Angstroms. The list of this dataset using the PDB chain identifiers is as follows:

1AAJ_ 01 _
1ADD_ 01 _
1AEP_ 01 _
1ALKA 02 A B
1AMG_ 01 _
1AMP_ 01 _
1AORA 02 A B
1AOZA 02 A B
1ARB_ 01 _
1ASZA 02 A B
1ATNA 02 A D
1AZCB 02 A B
1BABB 04 A B C D
1BBPA 04 A B C D
1BET_ 01 _
1BMTA 02 A B
1CAUA 02 A B
1CAUB 02 A B
1CCR_ 01 _
1CELA 01 A
1CEWI 01 I
1CFB_ 01 _
1CID_ 01 _
1CMCA 02 A B
1COLA 02 A B
1CPCA 04 A B K L
1CPCB 04 A B K L
1CRL_ 01 _
1CSN_ 01 _
1CTM_ 01 _
1CTN_ 01 _
1CHMA 02 A B
1DAAA 02 A B
1DHR_ 01 _
1DPRA 02 A B
1DSBA 01 A
1DYNA 02 A B
1EDE_ 01 _
1EDT_ 01 _
1EFT_ 01 _
1FBAA 04 A B C D
1FCDC 04 A B C D
1FKF_ 01 _
1FNC_ 01 _

1FRUB 06 A B C D E F
1GBS_ 01 _
1GHSA 02 A B
1GKY_ 01 _
1GMFA 02 A B
1GOF_ 01 _
1GPR_ 01 _
1HDCA 04 A B C D
1HEX_ 01 _
1HFC_ 01 _
1HMY_ 01 _
1HNF_ 01 _
1HSLA 02 A B
1HTMD 06 A B C D E F
1HUCB 02 A B
1HVD_ 01 _
1IAE_ 01 _
1IAG_ 01 _
1IGP_ 01 _
1KAB_ 01 _
1L92_ 01 _
1LBA_ 01 _
1LKI_ 01 _
1LPE_ 01 _
1LTSA 07 A C D E F G H
1LTSD 07 A C D E F G H
1MSAA 04 A B C D
1NAL1 04 1 2 3 4
1NAR_ 01 _
1NBAA 04 A B C D
1NDH_ 01 _
1NNT_ 01 _
1NPK_ 01 _
1PBE_ 01 _
1PBP_ 01 _
1PCRH 03 H L M
1PHP_ 01 _
1PHR_ 01 _
1PMY_ 01 _
1PNE_ 01 _
1POA_ 01 _
1POC_ 01 _
1PPI_ 01 _
1PRCC 04 C L M H
1PRCL 04 C L M H
1PRTC 12 A B C D E F G H I J K L
1PRTD 12 A B C D E F G H I J K L
1PSPA 02 A B
1RCB_ 01 _
1RIBA 02 A B
1RTP1 03 1 2 3
1RVAA 02 A B
1SACA 05 A B C D E
1SCUA 04 A B D E
1SCUB 04 A B D E
1SRYA 02 A B

1STD_ 01 _
1TADC 03 A B C
1TCA_ 01 _
1THV_ 01 _
1TIB_ 01 _
1TLCA 02 A B
1TLK_ 01 _
1TPH1 02 1 2
1TRKA 02 A B
1TSSA 03 A B C
1ULA_ 01 _
1VAAA 03 A B P
1XNB_ 01 _
1YPTB 02 A B
1YTBA 02 A B
2ACG_ 01 _
2BBKH 04 H L J M
2CPL_ 01 _
2CTC_ 01 _
2CHR_ 01 _
2END_ 01 _
2FCR_ 01 _
2GSTA 02 A B
2HNQ_ 01 _
2HPDA 02 A B
2KAUB 09 A B C D E F G H I
2KAUC 09 A B C D E F G H I
2LIGA 02 A B
2LIV_ 01 _
2MGE_ 01 _
2OHXA 02 A B
2PGD_ 02 _ A
2PIA_ 01 _
2RN2_ 01 _
2SAS_ 01 _
2SCPA 02 A B
2SIL_ 01 _
2SNV_ 01 _
3AAHA 04 A B C D
3MDDA 02 A B
4BLMA 02 A B
4ENL_ 01 _
4FXN_ 01 _
5P21_ 01 _
6FABL 02 L H
6TAA_ 01 _
8ACN_ 01 _

The syntaxis of the list is as follows:
XXXY NN A B C D E
where "XXX" is the PDB identifier of the protein, "Y" is the non-redundant chain ("_" means that the chain has no identifier) , "NN" is the number of protein chains considered in the calculation of the mean force potential and "A B C D E" is the specification of the

chain identifiers considered in the calculation.

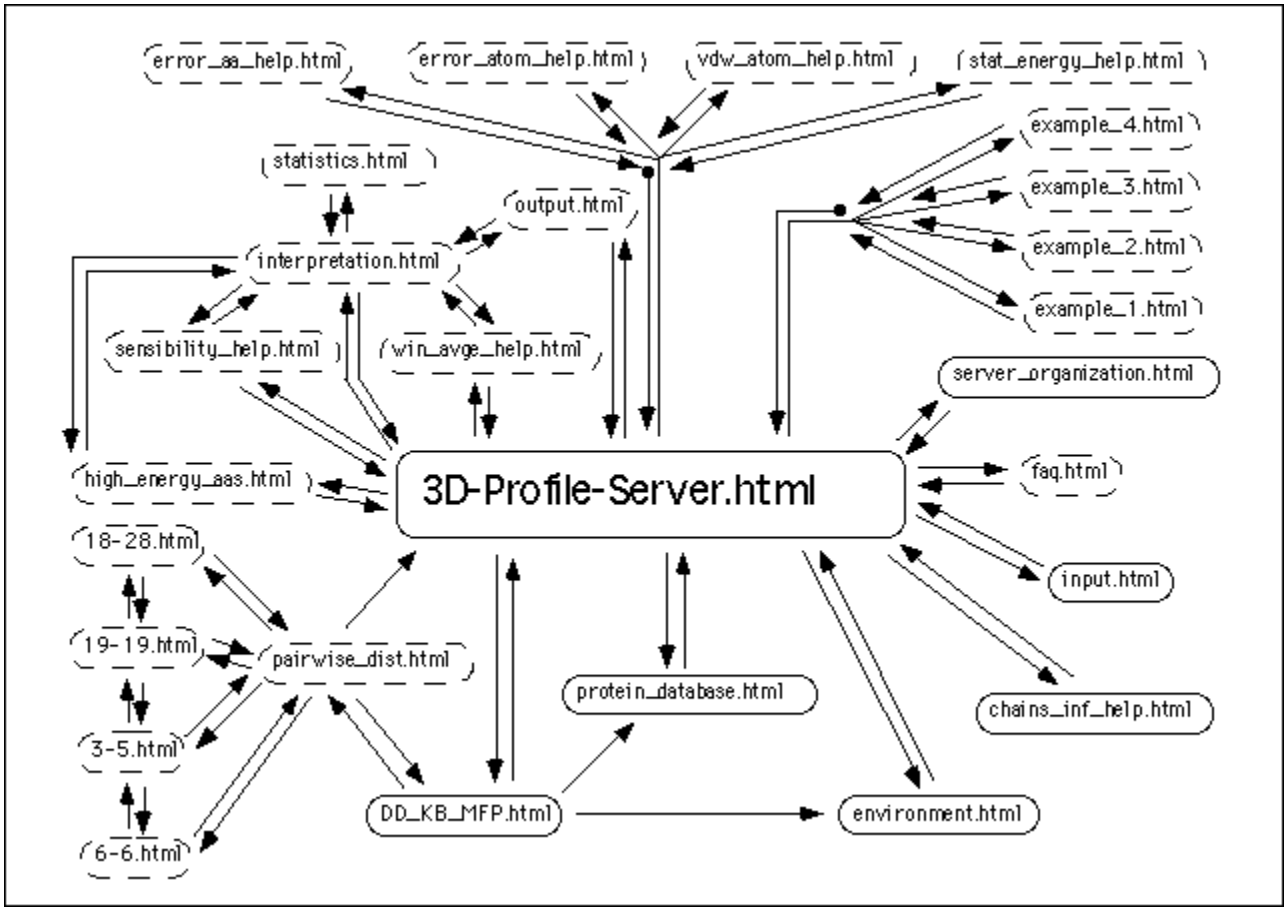
The calculation was performed on the non-redundant protein chain only (one per protein) but considering all the other heavy atoms of the other chains, if there are any.

Bibliography

- Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992) "Selection of representative protein data sets". *Protein Science* **1**, 409-417.
- Hobohm, U. & Sander, C. (1994) "Enlarged representative set of protein structures". *Protein Science* **3**, 522.

Server organization

This server is organized as shown in the figure below. Continuous line frames represent already functional html pages linked with the server and dashed line frames represent not yet available pages, but they will be implemented soon. The link directions are shown by the arrows.



How can I use ANOLEA ?

You must submit the Cartesian coordinates for each heavy atom of a protein (monomer or multimer). ANOLEA will calculate a non-local energy profile and you will obtain a list of the "high energy" amino acids for one chain of this protein, depending on the sensibility and the window average used to perform the calculations.

What is the usefulness of ANOLEA ?

The high energy amino acid list obtained after calculate the energy profile could represent some of the following features:

- errors in a model or in an experimentally solved protein structure. (example)
- potential interacting zones between subunits in a multimeric protein. (example)
- potential interacting zones between different proteins (example)

Window average help

As default, a moving window average of 5 residues is used. Using this value, a threshold of zero E/kT units is appropiated to define a high energy region. When changing the window average value, the user should be careful about the interpretations and the threshold defined. We provide the flexibility to fix the values for the window average and threshold. This is intended to be used carefully.
The user could want to see the energy profile using a window average of 1, after some regions have been identified as suspucious ones. The value of 1 in the window average parameter means that no average is used, thus each residue can be examinated independently of its neighbors. With this analysis, the residues that are mainly responsible for the high energy regions obtained can be pointed out.

Threshold help

The threshold value can be changed to any real number. As default, a value of zero is defined. In some cases, i.e. comparative models assessment, the user could be interested only in checking the existence of misalignments or regions that have very high energy scores. In that case, the threshold could be conveniently changed in such a way to display only those regions in the profile or in the graphic output.

Chain information help

Suppose that you want to submit a protein that contains two chains, named A and B. Then you have four possibilities to obtain an energy profile:

1.- To fill the chain information box with the next sequence of characters:
AB
With this information, the profile will be performed on chain **A**, but considering all the atom contacts with the chain **B**.

2.- To fill the chain information box with the next sequence of characters:
BA
With this information, the profile will be performed on chain **B**, but considering all the atom contacts with the chain **A**.

3.- To fill the chain information box with the next sequence of characters:
A
With this information, the profile will be performed only on chain **A**, without considering the atoms of chain **B**, even if you submit the atom coordinates of this chain.

4.- To fill the chain information box with the next sequence of characters:
B
With this information, the profile will be performed only on chain **B**, without considering the atoms of chain **A**, even if you submit the atom coordinates of this chain.

If you are submitting a unique protein chain without identifier, you must leave the chain information box empty.

If you are submitting a unique protein chain with identifier, you must write it in the chain information box.