The MESHI 9.17 energy terms and features

by

Chen Keasar

Department of Computer Science, Ben-Gurion University, Beer Sheva, Israel

Technical Report #15-06

October 2015

The MESHI 9.17 energy terms and features

Chen Keasar, Departments of Computer Science and Life Sciences,

Ben Gurion University of the Negev, chen.keasar@gmail.com

Introduction

MESHI (Kalisman *et al.*, 2005) is a protein modeling package designed to support the development of novel energy functions, optimization, and analysis methods. This document lists the current set of features provided by MESHI. The features are energy terms and other numerical values, which can be measured from a three dimensional protein model.

Many of the following features are currently in different developed stages. Many other features are obsolete, relicts of old less successful projects. Yet, they leave traces in publically available datasets so I provide them here for the sake of completion.

In what follows:

a. ***D*** represents a function whose ***d***erivative is also provided by MESHI. Typically these are energy terms used for force-driven simulations.
b. ***M*** represents a ***m***ultibody term, that is one in which the number of interacting atoms depends on protein size and/or geometry
c. ***C*** represents a ***c***ooperative term with non-additive contributions of its components.
d. ***O*** indicates an ***o***bsolete feature, apparently a dead end in the development process. These features are kept for backward compatibility (probably not for ever).
e. ***A*** indicates that the feature is ***a***ctive, that is focus of current intensive study.
f. ***T*** – A me***t***a-term that considers the distribution of another term.

Standard bonded terms

These terms are similar in spirit and functional form to related terms in legacy software (e.g., CHARMM). While not very informative (i.e. a geometrically perfect model is easy to make) these terms are essential in Cartesian space simulations, where they ensure the physical feasibility of the models.

1. Bond (***D***) – a quadratic penalty to deviation of bond lengths from ideal values.
2. Angle (***D***) – a quadratic penalty to deviation of bond angles from ideal values.
3. Plane (***D***) - a quadratic penalty to deviation of planar torsion angles (e.g., peptide bond) from planarity.
4. OutOfPlane (***D***) - a quadratic penalty to deviation of improper angles, associated with chiral atoms, from their ideal values.

Torsion angle terms

These terms quantify the compatibility of torsion angles with the Ramachandran plot and rotamer preferences.

5. RamachandranSidechain (***D***) – knowledge based term that considers all the torsion angles of a residue (Amir *et al.*, 2008). The reference state is an even distribution.
6. RamachSTD (***MTC***) – The standard deviation of per-residue RamachandranSidechain energy (5). This term tends to correlate with decoy quality, apparently because many decoys include both close-to-perfect secondary structure geometry and badly shaped coil regions.
7. CooperativeZRamachandranSidechain (***MTCOD***) & cooperativeZstdRamachandranSidechain (***MTCOD***) – These terms penalize deviation from the expected per-residue mean and standard deviation of RamachandranSidechain energy (5). They do not seem to be of much use, at least in the current functional form.
8. Propensity (***D***) – knowledge based term that considers only the backbone residue (Amir *et al.*, 2008). The reference state is the average distribution over all the residue types.
9. RamachandranCore (RamachandranSideChainCore) (***D***) – Approximates the allowed regions of Ramachandran plot by flat rectangle wells. Note that the in older versions this term has a misleading name, which suggests that it considers side-chains.
10. FlatRamachEnergy (***DO***) – An earlier version of RamachandranSidechain (5).

Hydrogen bond terms

Hydrogen bond networks are a hallmark of protein structures. We have studied them in two parallel directions and ended up with two sets of terms. We hope to consolidate them at some stage.

11. HydrogenBond (***D***) – A simple term that considers the distance between backbone hydrogen and oxygen atoms (Levy-Moonshine *et al.*, 2009).
12. HydrogenBondsAnglesHOC (***D***) & hydrogenBondsAnglesHOC (***D***) – These terms penalize backbone hydrogen bonds whose corresponding angles are too small (Levy-Moonshine *et al.*, 2009).
13. SolvationHB (***DA***) – A more complex hydrogen bond term, which is based on (Dahiyat and Mayo, 1996) . It considers both bond distances and angles, and is applied to all hydrogen bonds in the protein. The name indicates that it is calculated by the Solvation energy function (see below 36) as a by-product.
14. HydrogenBondPairs (***DCA***) – Built on top of the hydrogenBond (11), this cooperative term rewards or penalizes pairs of backbone hydrogen bonds depending on their frequency in native structures (Levy-Moonshine *et al.*, 2009).

Radius of Gyration

These features quantify two observations about the Radius of Gyration (RG) in native protein structures:

a. There is close to linear relation between the log of the protein length (**logL**) and the log of the radius of gyration (**logRG**).
b. There are predictable relations between the RGs of distinct subsets of residues: hydrophobic vs. polar and helices and beta vs. coil segments.

While useful these features seem to be highly correlated to, and noisier than, related contact based features. Thus, we consider them obsolete.

15. N_RGhSS (***MCDO***) – Deviation of logRG of hydrophobic side-chains in secondary structure elements from the value expected according to logL.
16. hSS (***MCDO***) – logRG of hydrophobic side-chains in secondary structure elements.
17. bSS (***MCDO***) – logRG of Cα atoms in secondary structure elements.
18. cSS (***MCDO***) – logRG of polar atoms in secondary structure elements.
19. hSShCoil (***MCDO***) – the ratio between logRG of hydrophobic side-chains in secondary structure elements and logRG of hydrophobic side-chains in coil segments (tends to be lower than one in native structures).
20. EhSShCoil (***MCDO***) – penalize severe deviations from expected value of the above ratio.
21. hSSbSS (***MCDO***) – the ratio between logRG of hydrophobic side-chains in secondary structure elements and logRG of their Cα atoms (tends to be lower than one in native structures).
22. EhSSbSS (***MCDO***) – penalize severe deviations from expected value of the above ratio.
23. hSSbCoil (***MCDO***) – the ratio between logRG of hydrophobic side-chains in secondary structure elements and logRG of Cα atoms in coil segments (tends to be lower than one in native structures).
24. EhSSbCoil (***MCDO***) – penalize severe deviations from expected value of the above ratio.
25. hSScSS (***MCDO***) – the ratio between logRG of hydrophobic side-chains in secondary structure elements and logRG of polar side-chains in secondary structure elements (tends to be lower than one in native structures).
26. EhSScSS (***MCDO***) – penalize severe deviations from expected value of the above ratio.
27. hSScCoil (***MCDO***) – the ratio between logRG of hydrophobic side-chains in secondary structure elements and logRG of polar side-chains in coil segments (tends to be lower than one in native structures).
28. EhSScCoil (***MCDO***) – penalize severe deviations from expected value of the above ratio.
29. Rg (***MCDO***) – sum over 20, 22, 24, 26, and 28.

Summa and Levitt pairwise PMF

We use the a truncated version of the pairwise potential of Summa and Levitt (Summa and Levitt, 2007) as our basic non-bonded energy term during conformational searches. The potential is truncated at 5.5 Å to reduce computational price. The energy value may be decomposed to atom contributions, and we noted that after energy optimization the distributions of these energies differ considerably from what is observed in native structures. Generally speaking, the energies are lower than expected.

30. AtomicPairwisePMFSumma (***D***) – a truncated (at 5.5 Å) version of (Summa and Levitt, 2007).
31. SummaStd (***MTC***) – The standard deviation of the per-atom AtomicPairwisePMFSumma energies (30). Correlates with decoy quality as high values indicate loosely packed decoys with highly packed cores.

32. CooperativeZSumma (***MTDCA***) – A term that penalizes deviation from the expected distribution of per-atom AtomicPairwisePMFSumma energies (30).
33. CooperativeSummaPolar (***MTDCA***) – The contribution of polar side-chain atoms to CooperativeZSumma (32).
34. CooperativeSummaNonPolar (***MTDCA***) – The contribution of non-polar atoms to CooperativeZSumma (32).
35. CooperativeSummaPolarNN_OO (***MTDCA***) – The contribution of backbone hydrogen bonds to CooperativeZSumma (32).
36. CooperativeSummaPolarBb (***MTDCA***) – The contribution of polar backbone atoms to CooperativeZSumma (32), excluding hydrogen bonds.
37. CooperativeZstdSumma ,CooperativeStdSummaPolar, CooperativeStdSummaNonPolar, CooperativeStdSummaPolarNN_OO, CooperativeStdSummaPolarBb (***MTDCO***) – an obsolete older version of features 32-36.

Atom environment energy term

Inspired by (Delarue and Koehl, 1995; Summa *et al.*, 2005), but with a different functional forms and based on a far larger database of native structures. This energy term considers two polar atoms as neighbors if they are hydrogen bonded. Any other pair of atoms is considered neighboring if their distance is 5.5 Å or lower. We define the atomic environment as a pair (CNC, PNC), where CNC stands for a Carbon Neighbors Count, and PNC stands for Polar Neighbor count.

38. solvationEnergy (***MDCA***) – An energy term derived from the distributions of atom environments in a dataset of native structures, and an "average-atom" reference state (Samudrala and Moult, 1998).
39. solvationSCpolar (***MDCA***) – The contribution of polar sidechain atoms to solvationEnergy (38).
40. solvationBBcarbon (***MDCA***) – The contribution of backbone carbon atoms to solvationEnergy (38).
41. solvationBBpolar (***MDCA***) – The contribution of polar backbone atoms to solvationEnergy (38).
42. solvationSCcarbon (***MDCA***) – The contribution of carbon sidechain atoms to solvationEnergy (38).
43. solvationBuriedHB (***MCA***) – The number of buried hydrogen bonds
44. solvationSTD (***MCA***) – The standard deviation of atom solvationEnergy (38) energies. It is correlated with decoy quality as bad decoys often have well packed cores and/or highly exposed charged atoms along with badly packed regions and buried polar atoms, which are not hydrogen bonded.
45. solvationEntropy (***MCA***) – similar in spirit to solvationSTD (44) but tries to capture the shape of the whole energy distribution. Seems less useful in the current formulation.
46. bbPolarN, bbCarbonN, scPolarN, scCarbonN – the number of atoms corresponding to features 39-42.

Simple contact features

These features are by-products of some abandoned project, but they turn up useful so we keep them. Contacts are defined here as pairs of atom whose distance is below some threshold. The thresholds of these features are larger than the rather tight (5.5 Å) threshold used for environment calculations (38), which

result in higher computational requirements. Thus, we do not use them for iterative force driven simulations and do not bother to derive them.

47. Contacts8 – Average number of Cα contacts with 8Å threshold.
48. contacts11 – Average number of Cα contacts with 11Å threshold.
49. Contacts15 – Average number of Cα contacts with 15Å threshold.
50. contactsHr – Average all atoms contacts with 6Å threshold
51. ConservationContacts8, conservationRgRatio, conservation_H_RgRatio, conservationContacts11, conservationContacts15 – Obsolete features

Contact based features

Unlike the simple features above these features are knowledge based and refer to characteristic distributions in native structures. They used the same contact definitions as above and are not derivable due to the same reason.

52. Contacts12 (***MA***) – The average number of 12Å contacts hydrophobic side-chain atoms in secondary structure segments correlates well with the log of their number. Thus, the average number of contacts in a native structure can be predicted from its atom composition. This feature compares (by Z-score) the expected and observed average.
53. Contacts14 (***MA***) – The same as above (52) with a 14 Å threshold.
54. Contacts14Core (***MA***) - In native structures, the number of 14 Å contacts per hydrophobic atom in secondary structure elements is rather evenly distributed. This feature approximate deviation from this distribution by providing the Mahalanobis distance of length normalized average number of 14A contacts of all hydrophobic side-chains in SS and the average number of contacts among the most compact atoms (core).
55. SASAratio (***MA***) - The laziest feature in MESHI. Mahalanobis distance of length normalized average SASA (as measured by DSSP (Kabsch and Sander, 1983))of all hydrophobic side-chains in SS and the polar ones.

Miscellaneous

56. TetherEnergy, tetherAll (***DA***) - energy terms that restrain the model (or subset of it) to its initial position in force driven simulations.
57. samudralaEnergy, samudralaEnergy1 – Tow identical copies (one is removed in later versions) of an in-house implementation of (Samudrala and Moult, 1998).
58. One – 42^0
59. SecondaryStructureFraction – the ratio between the number of helix and beta residues and the number of coil residues.
60. DistanceConstraints – distance constraints imposed during the simulation.
61. Energy – a linear combination of many of the above values. Coefficients were determined by trial and error.

References

Amir,E.-A.D. *et al.* (2008) Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins: Structure, Function, and Bioinformatics*, **72**, 62–73.

Dahiyat,B.I. and Mayo,S.L. (1996) Protein design automation. *Protein Science*, **5**, 895–903.

Delarue,M. and Koehl,P. (1995) Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J. Mol. Biol.*, **249**, 675–690.

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

Kalisman,N. *et al.* (2005) MESHI: a new library of Java classes for molecular modeling. *Bioinformatics*, **21**, 3931–3932.

Levy-Moonshine,A. *et al.* (2009) Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. *Bioinformatics*, **25**, 2639–2645.

Samudrala,R. and Moult,J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, **275**, 895–916.

Summa,C.M. *et al.* (2005) An Atomic Environment Potential for use in Protein Structure Prediction. *Journal of Molecular Biology*, **352**, 986–1001.

Summa,C.M. and Levitt,M. (2007) Near-native structure refinement using in vacuo energy minimization. *PNAS*, **104**, 3177–3182.