Correlation Functions and Property Codes of Different Proteomes

Olaf Weiss*and Hanspeter Herzel

Innovationskolleg Theoretische Biologie, Humboldt University Berlin, Invalidenstr. 43,

D–10115 Berlin, Germany

The successful realization of genome sequencing projects has produced lots species specific sequence data. Usually, genomes are compared gene by gene applying the standard tools such as pairwise alignments. However, it is also possible to compare genomes by global measures such as overall G+C-content or its periodicities [1]. We intend to make a comparison on a global level based on protein sequences.

Using correlation functions in protein sequences [2] and principal component analysis we have developed an iterative procedure that finds the property code that gives the strongest correlation in a given proteome. Like genetic algorithms, this iteration can be understood as a dynamical system. The dynamics of the system is determined by the sequence data used as input. Varying transient dynamics, and unique as well as multiple fixed points are observed. For instance, the iteration in the yeast proteome can exhibit two fixed points corresponding to a hydrophobicity-like property code and a property similar to α-helix propensity. The initial conditions of the iteration determine which one of the fixed points is reached.

The fixed points of such systems can be identified with classifications or "properties" of the amino acids. The dynamics as well as the result of the iteration differ significantly for different kinds of sequence data used. We discuss our results for the superfamily set used in [2] and for the proteomes of E. coli, M.jannaschii and yeast.

References

[1] H. Herzel, O. Weiss, & E.N. Trifonov, *10-11 bp periodicities in complete genomes reflect protein structure and DNA folding*, Bioinformatics (1999) **15**(1).

[2] O. Weiss & H. Herzel, *Correlations in protein sequences and property codes*, J. theor. Biol. (1998) **190**.

*e-mail: o.weiss@biologie.hu-berlin.de