**Christoph Adami "Information Theory in Biology" Physics of Life Reviews (2004)**

**Summary**

Adami explains Shannon’s information theory and then applies it to sequences, proteins, and drug development. Information (I) is entropy (H) that is shared between two systems ( I(X:Y) = H(X) + H(Y) – H(XY) ). Information theory can be applied to comparisons between members of a population to determine which elements of a sequence are information rich, polymorphic or neutral. An information analysis between two elements can be used find co-varyiation. This co-varyiation might help to predict interactions and produce better base alignments (through minimization).

**Glossary**

Shannon paper on information theory (1948 Bell Systems - "Theory of Communication)

Entropy in Shannon an observer's uncertainty about the identity of an object

H(x) Entropy = -sum across all outcomes of (probability)\*log(probability)

H(x)max Highest possible entropy occurs if all outscomes are equaprobable and

H(x) = log(N) where N is the length

# of ?'s H(x)max base is the number of yes/no questions needed to identify x

I(x) Information = H(x)max - H(x)

I(X:Y) Information of x on y = H(x) + H(y) - H(xy)

I(X|Y) Conditional Entropy = H(XY) - H(Y)

I(X:Y) H(X) - H(X|Y)

Epistatic Correlations bp between two distant nucleotides as in base pairing in tRNA

Information in DNA 2L - H where H = entropy across the positions of the sequence \*if 2 bits/base

degrees of freedom variables that can change

coarse graining continous becomes discrete due to the insturments used to measure it.

Information entropy that is shared between two systems

Statistical ensemble identically prepared sets (like homologous sequences)

**Claims**

Entropy of DNA approx. 2 bits/base (Schmitt Herzel 1997)

optimal code cannot be further compressed (Cover Thomas 1991)

protein compression only 1% is possible (Weiss 2000 J of theoretical biology)

*Myco. pneumonia* 89% info in soft tissue host (Dandekov 2000 Nucleic Acids Research)

polymorphism binds multiple partners and 0 < I(x) < 2 bits (Adami - Physics of Life Rev. 2004)

I(DNA) Maps inter-sequence communication. Like co-varient binding in a tRNA molecule (Adali)

I(proteins) Co-variation determines paired interactions, low-variation determines information, partial information determines polymorphisms (Adali)

I(cells) Used to correlate DNA to binding protiens, proteins to proteins and so on

I(drugs) could be used to find polymorphism and co-varying mutations

**Questions:**

How would one measure the complexity of an organism (see the C-paradox)?

What is the compresibility of DNA (Adami says it is low, without clear citation)?

Does stomach acid destroy viruses? Are microbial populations in the gut in viral communication with the world?

How is the information content (89%) of *Myco. P.* determined in its host? (by sequence alignement?)

How is physical and functional complexity measured? (adami 2002 bioessays; 2000 PNAS; Physics D.)

**Inspirations:**

A map of information between elements of the genome could be used to find interactions and help to align sequeces (align by maximum information)

We could add an analysis of modifications (ie. methylation) to and information theory analysis of DNA.

DNA information is being actualized all the time in the way of RNA and viral copies.

**Programs:**

To calculate the entropy at one base for a set of nucleotide probabilities <http://localhost/Archives/Entropy_of_one_base.py>