The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2058-S-13

Pseudochaos and Stable-Chaos in Statistical Mechanics and Quantum Physics

21 - 25 September 2009

Applications of Mutual Information: from Cardiology to Molecular Genetics

P. Grassberger
University of Calgary
Canada

Stable or Non-Extensive Chaos?

Trieste,
September 23, 2009

chaos $\leftrightarrow$ “small” perturbation $\leftrightarrow$ metric (distance)

Finite-dimensional systems: all metrics (Euclid, max-norm, ...) are equivalent

$$\text{“small”} = \text{“small”}$$

Infinite-dimensional vector spaces: inequivalent norms:

- sup norm:
$$||\mathbf{x}||_{\text{sup}} = \sup_i |x_i|$$
- L^p, l^p norms:
$$||\mathbf{x}||_p = [\sum_i |x_i|^p]^{1/p}$$
- $$||\mathbf{x}||_{\text{loc}} = \sum_{i=-\infty}^{\infty} \frac{|x_i|}{2^{|i|}}$$

Last example: not translation invariant, i.e. if $(T\mathbf{x})_i = x_{i+1}$, then

$$||T\mathbf{x}||_{\text{loc}} \neq ||\mathbf{x}||_{\text{loc}}.$$

Nevertheless, it is a good norm (positive, symmetric, triangle inequal., ...)

Conjecture:

All examples of “stable” chaos are bona fide chaotic systems, if

- limit system size $\rightarrow \infty$ is taken first
  (“transients” become strictly stationary invariant states)
- A suitably defined “local” norm like $||\mathbf{x}||_{\mathrm{loc}}$ is used.

Comments:

1) $\mathrm{Lim}_N \to \infty$ is as natural as taking infinitely many digits in, say, the Roessler equations.

2) Some constructions familiar from finite dimensional chaos pass through:

a) Lyapunov exponents = velocitites of left-, right-moving perturbations

b) KS entropy of finite subsystem of ∞ system: = Shannon information needed per time step to specify state

c) (information) Dimension density = information needed to specify state at one site, at fixed time

d) Pesin-Ledrappier-Young formula (S. Wolfram, Physica D, 1984):

$$\text{KS entropy} \leq \sum\nolimits_+ \text{Lyapunov exp.} \times \text{dimension density}$$

Main difference to “unstable” chaos:

In ordinary (“unstable”) chaos, KS entropy is extensive (finite entropy density, produced locally)

In “stable” chaos: uncertainty is carried in from outside
→ KS entropy is proportional to or smaller than surface

1 dimension: entropy = const.

“non-extensive chaos”

New insight besides new words, new results, ...:

???

Therefore .....

Applications of Mutual Information: From Cardiology to Molecular Genetics

FZ Jülich Group:

A. Kraskov	*Univ. College, London*
H. Stögbauer	*Caltech*
R. Andrzejak	*Univ. Pompeu Fabra, Barcelona*
S. Astakhov	*Unique ICS/KWE, Saratov, Russia*

Calgary Group:
D.V. Foster, O. Penner, M. Paczuski

Trieste,
September 23, 2009

Outline:

- **What is a zebra?**
  Mutual Information: a universal indicator of similarity & independency
- **Intelligent reductionism:**
  decomposing a complex systems into independent parts
- **Slaves, masters & genes**
  Who influences whom? Does similarity indicate dependency?
- **Putting similar things together:**
  Constructing a tree of Life
- **Zipping & BLASTing:**
  Another way of finding similarities between DNA or amino acid sequences

What is a zebra?

What is a zebra?

a horse with stripes!

If we have two very similar objects, & we know already one of them

→ little additional information is needed to specify the other!

$H(X)$ = information needed to specify object X
$H(Y)$ = information needed to specify object Y
$H(XY)$ = information needed to specify both X and Y
$H(Y|X)$ = additional information needed to specify Y, if we have already a complete description of X
$= H(XY) - H(X)$

Mutual Information between X & Y:

$$MI(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(XY)$$

= information about X which is also useful for describing Y & vice versa

$$MI(X;Y) \geq 0,$$

$= 0$ only, when X, Y completely independent

What means $H(X)$ really?

How can we measure / estimate it?

1) "Algorithmic Information Theory" [Kolmogorov, Chaitin, ...]
choose a computer (e.g. PC, mainframe, ...);
choose and operating system (LINUX, Windows, Apple, ...);
choose a compression algorithm (zip, gzip, bzip2, ...)

X = file
$H(X)$ = length of compressed file of best possible compression algorithm
E.g. X = Romeo & Juliet (project Gutenberg): 159789 bytes
after compression with gzip: 54208 bytes
after compression with bzip2: 47572 bytes

→ $H(\text{Romeo \& Juliet}) \leq 47572$ bytes

Estimation of MI:
first compress X, Y seperately, add their compressed lengths,
then concatenate them to XY & compress also XY,
subtract the two results
NB: never exact, since no "best possible" compression algorithm !

2) Probabilistic Information Theory [Carnot, Clausius, Boltzmann, ... Shannon]

X, Y are random variables

(x_i, y_i) are realizations $(i = 1, 2, \ldots N)$

→ estimate probabilities!

→ calculate log-likelihoods for observing x_i, y_i

$$H(X) = -\sum_x p_x \log p_x$$

= average negative log-likelihood.

Thm.: If $\{x_i;\ \ i = 1 \ldots N\}$ are random realizations of X, then $\frac{1}{N}\times$ {algor. inform. for describing all of them} $\to H(X)$ as $N \to \infty$ with probability 1.

Reductionism = dealing with a problem by decomposing it into “its parts”

- **Frequent criticism:** a human being cannot be meaningful dealt with by describing it in terms of atoms & quarks
- **Reply:** “its parts” is too naive & is not meant literally by any intelligent reductionist;

A complex object has so many possible decompositions that “its” parts really should read “our choice of its” parts

Better definition:

Intelligent Reductionism = dealing with a problem by decomposing it into adequate parts

Examples:

- Fourier decomposition
- Principle component decomposition
- **Independent component decomposition**
- **Least dependent component decomposition**

$\mathbf{X}$ = multivariate signal, e.g.:

- acoustic signal received by party guest at a party
- radio signal obtained by a cellular phone transmission station
- EEG (brain) signal
- ECG (heart) signal

Decompose **X** as well as possible into independent components!

Measure of independence: MI

Assume that $\mathbf{X}(t)$ is linear mixture of sources $\mathbf{S}(t)$ with constant (time independent) “mixing matrix $\mathbf{M}$:

$$\mathbf{X}(t) = \mathbf{M}\mathbf{S}(t)$$

Task: given $\mathbf{X}(t)$, find $\mathbf{M}$ and $\mathbf{S}(t)$ so that components $s_i(t)$ have minimal MI!

NB: PCA corresponds to minimal *linear* dependencies between $s_i(t)$

Special example: ECG of pregnant woman

(mother's ECG) + (fetal ECG) + noise

Task: extract clean fetal ECG!

H. Stögbauer, A. Kraskov, S.A. Astakhov, and P.G., PRE 2004:

- Use novel MI estimator (A. Kraskov, H. Stoegbauer, P.G., PRE 2004)
- Allow also for non-instantaneous (delayed) mixing
- Minimize also delayed MI with same delays
- Lump together components which cannot be made independent

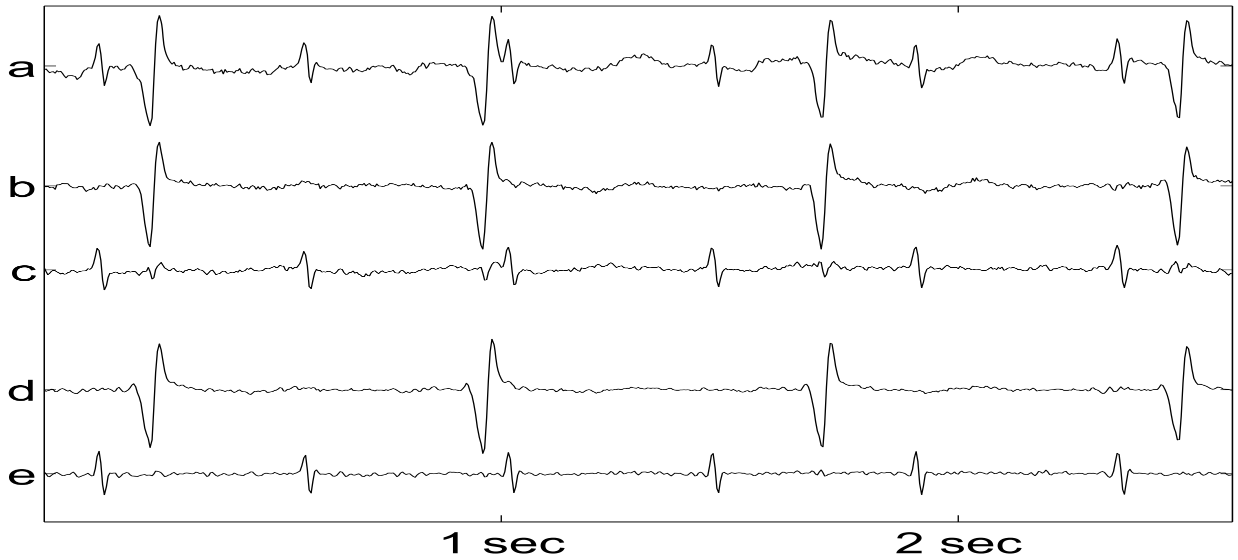

Figure 1: Short segment from the original ECG (a), of the mother and fetus contributions estimated without delay embedding (b,c), and of the two contributions estimated with delay embedding (d,e).

Who influences whom?

Slaves, masters & genes

“Wie der Herr, so es G’scherr”

[German proverb: “as is the master, so is the servant”]

- Can one turn this around?
  Is it true that A controls B, if A, B are similar?
- Can one use this for finding out who controls whom?
- Can one apply this to improve regulatory gene networks by means of large scale gene expression data?
- Gives MI better results than linear similarity measures (Pearson, Spearman)?

Problems:

- MI is symmetric:
  no distinction between $A \to B$, $B \to A$.
- If a master has two slaves, then they are also similar, but they don't control each other
- In gene regulation networks: often one gene needs two or more transcription factors, i.e. pairwise similarity can't be enough

Helpful:

- causality arrow = time arrow (if no delays in observation)
- Change from **passive observation** to **active experimentation** (gene knock-out experiments)
- If $MI(A;B) > 0,\ MI(B;C) > 0,\ \ MI(A;C) = 0$: then $A \rightarrow B \leftarrow C$
  is more likely than other three possibilities ($A \rightarrow B \rightarrow C$ etc.)
- Same new algorithm for MI estimation works also for generalized MI between > 2 genes
  (A. Kraskov, H. Stoegbauer, P.G., PRE 2004;
  S. Frenzel and B. Pompe, PRL 99, 2007)

MI gives more information than linear correlations:

K. Basso et al., Nature Genetics 2005:
Human blood cells (healthy & cancer),
12600 different gene loci
300 different B-cells

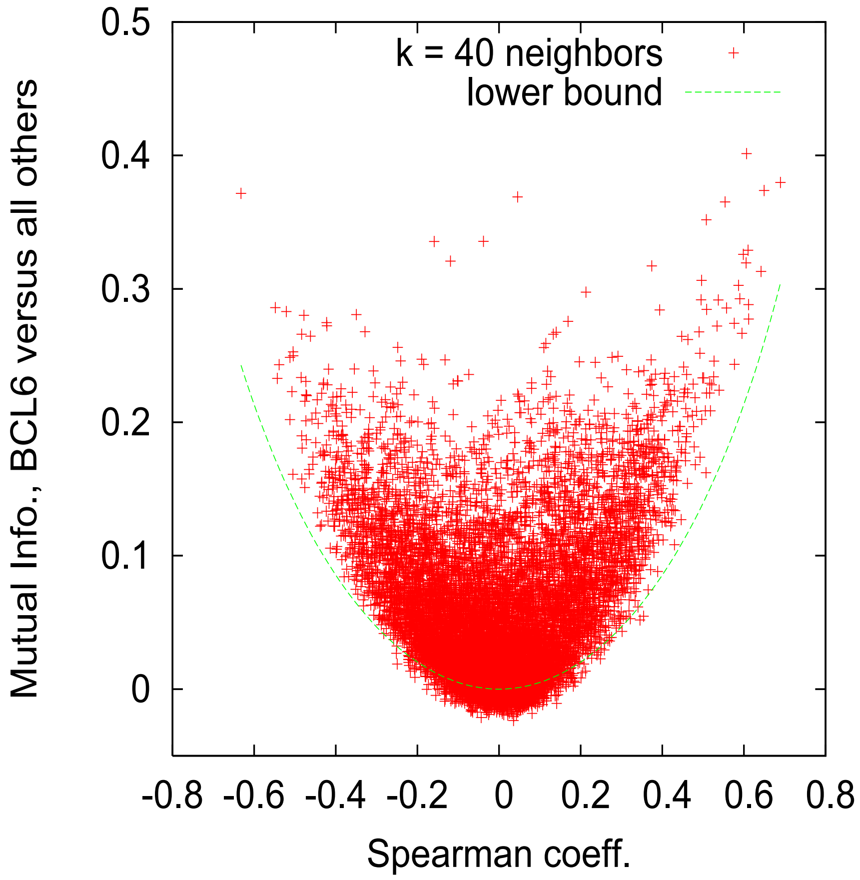

Figure 2: Mutual informations between gene BCL6 and all other 12500 genes, plotted versus Spearman coefficient. The green dashed line gives the theoretical lower bound. The spread of points below this curve gives a rough error estimate. Points high above it indicate nonlinear dependencies between the two genes.

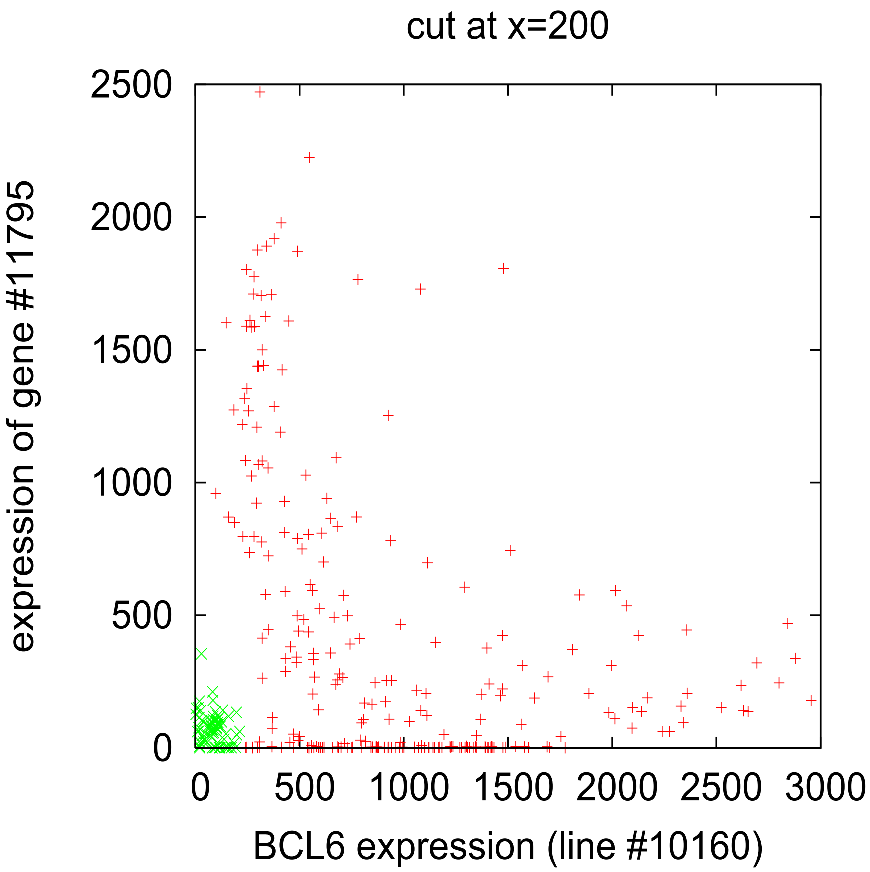

Figure 3: Expressions over all 300 cell types of genes BCL6 and gene # 10795, the latter corresponding to one of the four high & central points in the previous figure.

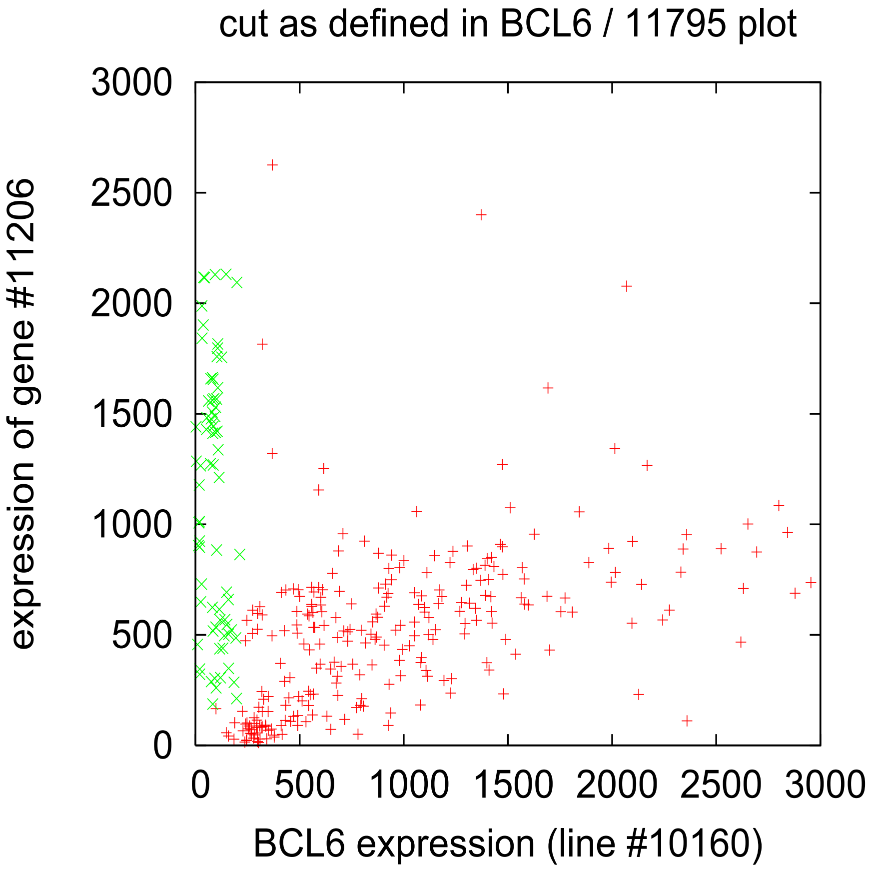

Figure 4: Same as previous, but for another of the four high & central points.

E. coli:

Large fraction of pairs shows significant $MI > 0$

Strongest MI between genes in same operon:
co-regulation more important than active/passive regulation

Co-regulated gene pairs have also large positive Pearson coefficient
→ transcription factors/regulated genes are enhanced in pairs where $MI >$ lower bound, Pearson coeff. is negative

Using conditional MI should be better than “data processing inequality” (ARACNE) in eliminating fake (indirect) interactions, but so far results unclear

- Data Processing Inequality (DPI): if $A \leftrightarrow B \leftrightarrow C$ without direct $A \leftrightarrow C$, then

  $$MI(A;C) \leq MI(A;B), MI(B;C)$$

  .

- ARACNE (Califano *et al.* 2005): reverse argument, i.e.
  if $MI(A;C) \leq MI(A;B), MI(B;C)$ then no direct $A \leftrightarrow C$.

- If no direct $A \leftrightarrow C$, then

$$MI(A;C|B) = 0$$

Not all MI decrease when conditioned!

Assume that neither X nor Y can regulate Z, but together they can:

$$\begin{aligned} & MI(X;Z) = MI(Y;Z) = 0, \;\; MI((XY);Z) > 0, \\ \rightarrow \quad & MI(X;Y;Z) = MIMI(X;Z) + MI(Y;Z) - MI((XY);Z) < 0. \end{aligned}$$

Can be checked for special triples (D. Anastasiou *et al.*, 2008), but test *all* possible triples (quadruples, ...)??

Constructing a Tree of Life for Animals:

All animals: mitochondrial DNA has $\approx 16,000$ base pairs, relatively similar

→ use mtDNA similarities to build phylogenetic trees!

Benedetto, Loreto, M. Li, Vitanyi, ...: use (algorithmic) MI!

1. estimate all pairwise MI's by concatenating & zipping
2. turn this matrix into matrix of *distances* ("universal compression distance")
3. use this with your favorite hierarchical clustering algorithm to build phylogenetic tree

It works!

Also:

Classification of:
  texts,
  music,
  languages,
  snowball letters,
  copied homework asignments,
attribution of texts to authors,
family trees,
  ...

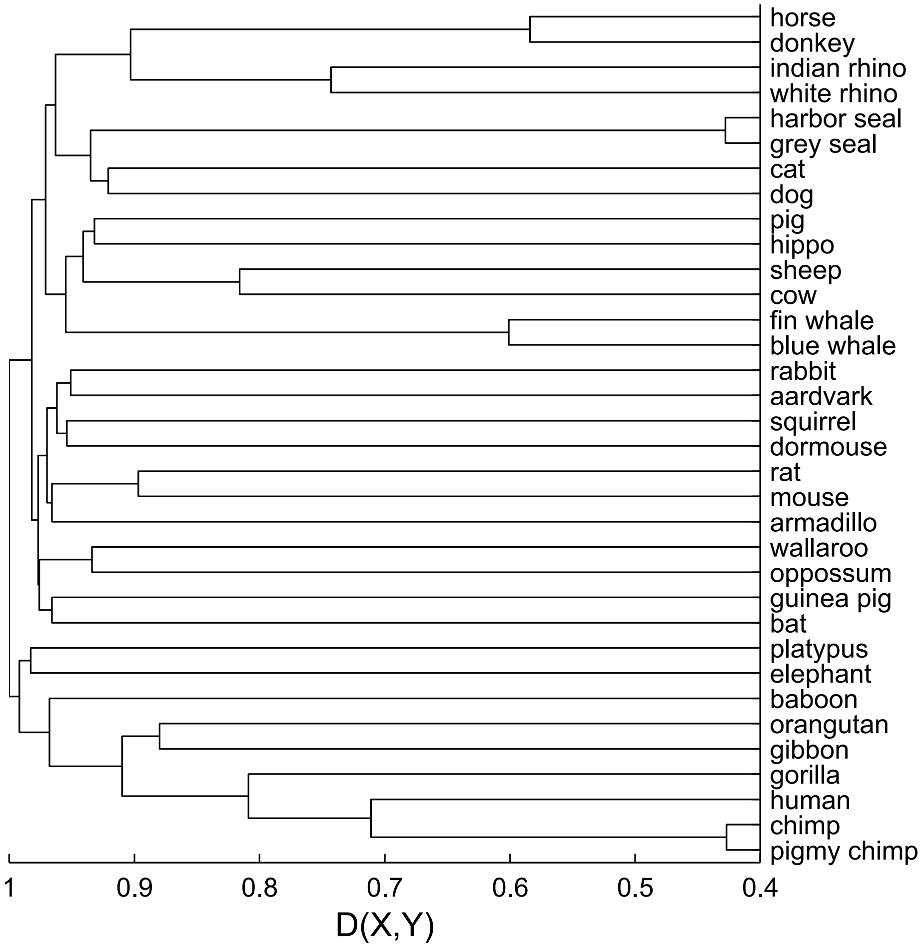

Figure 5: Phylogenetic tree for 34 mammals, based on mitochondrial DNA compressed with bzip2. The heights of nodes are the distances between the joining daughter clusters.

Does this mean that existing data compression algorithms are perfect?

No!

All general purpose compression algorithms and most dedicated DNA compression algorithms do badly for DNA

Exception: “XM” (“eXpert Model”), Cao *et al.* 2007

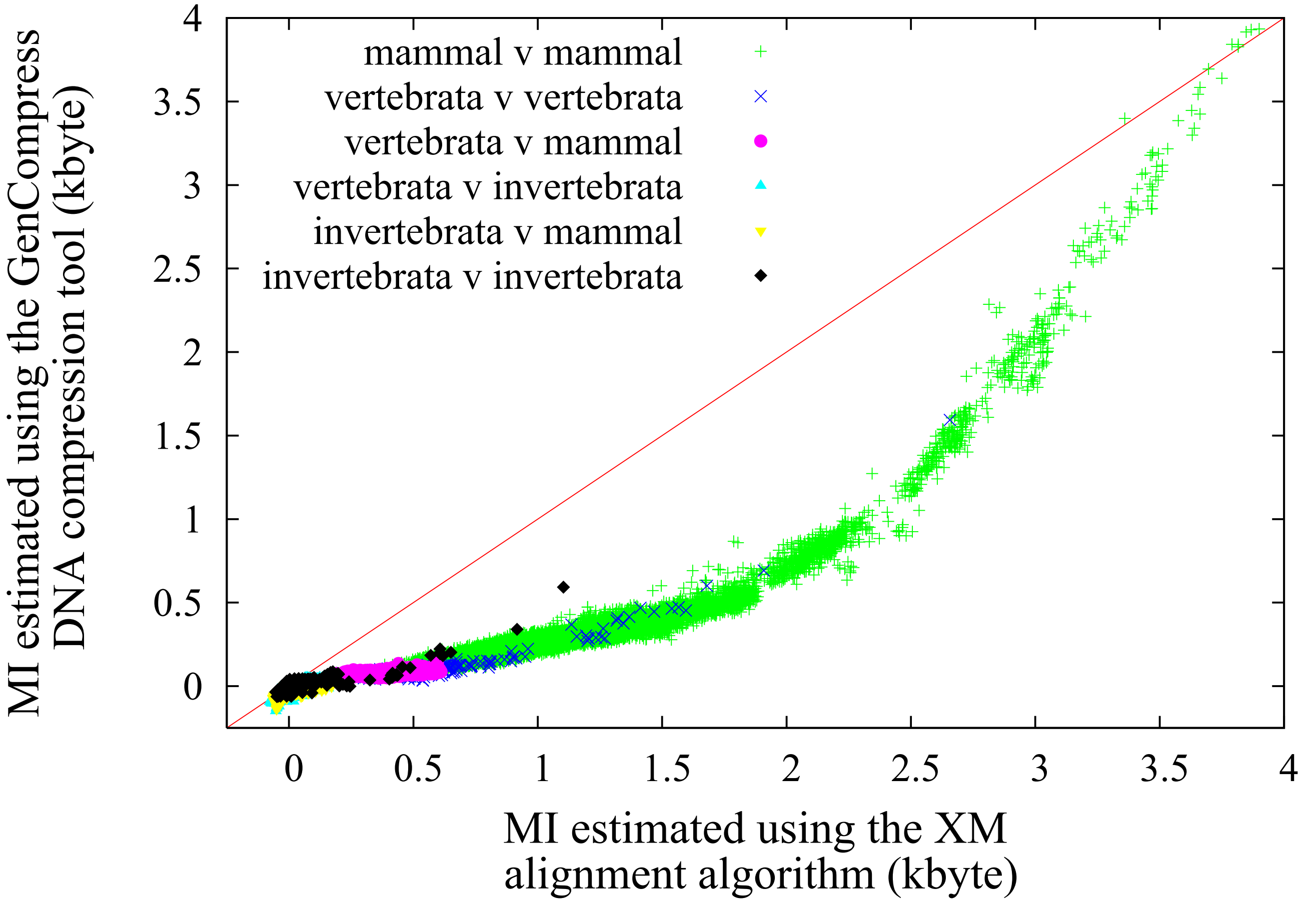

Figure 6: Compression-based MI estimates for complete mitochondrial DNA. x-axis: using XM; y-axis: using GenCompress.

Alternative estimation of MI between DNA sequences:

Sequence Alignment (“BLASTing”)

Global pairwise alignment:

insert blanks at suitabled places in both sequences, so that agreements at same position are maximized (“score” is maximized)

If good pairwise global alignment is given, then estimation of MI is easy!
(O. Penner, M. Paczuski, P. G., arXiv (2009))
In this estimate one needs also a compression algorithm, but its efficiency is much less important

Both estimates agree perfectly for similar species (e.g. primate / primate)

Alignment-based estimates slightly better for intermediate distances

XM-based estimates better for distant species (alignment breaks down)

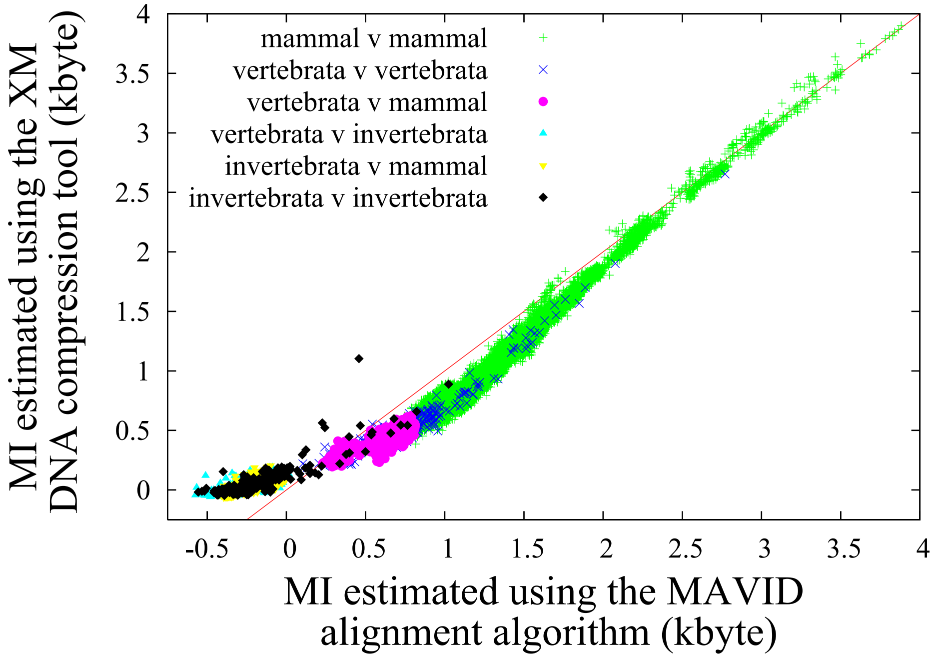

Figure 7: Zipping-based MI estimates (using XM algorithm) versus blasting-based MI estimates (using MAVID alignment algorithm).

Conclusions

- MI is a universal measure for interdependence with numerous applications (many here not even touched!)
- Estimating MI is in general not easy.
- Details depend much on whether algorithmic of probabilistic information theory is more appropriate
- In both cases, use of MI allows a conceptually very simple clustering algorithm
- Clustering of sequences using compression works surprisingly well, given the huge imperfections of present compression algorithms
- independent component analysis with k-nearest neighbor MI estimates & clustering gives excellent results
- Comparing MI with linear (Spearman) correlations allows to detect nonlinear interdependences.

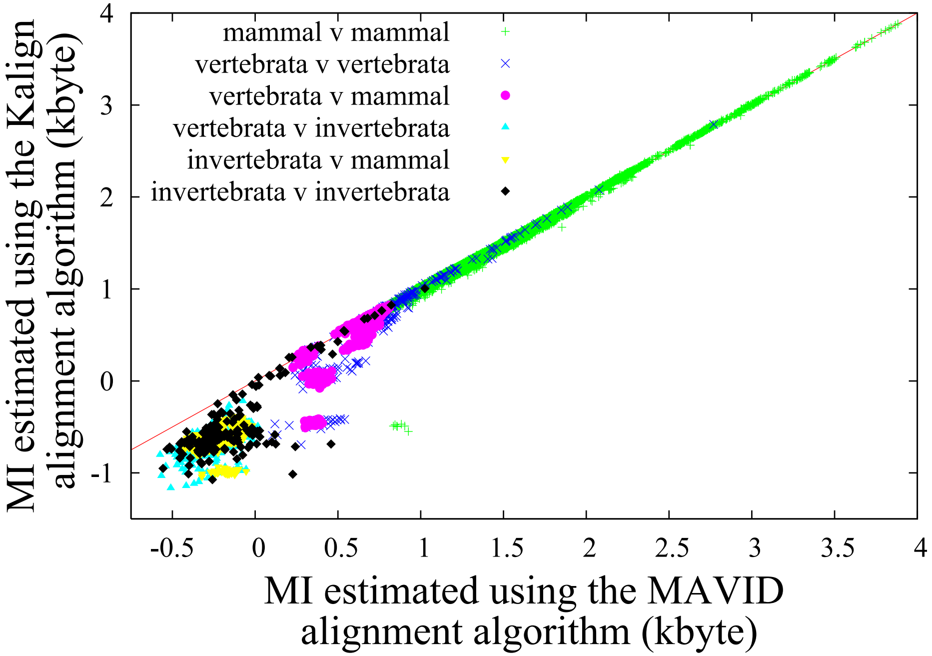

Figure 8: Blasting-based MI estimates using KALIGN versus blasting-based MI estimates using MAVID.

- Conditional MI can improve network inference
- Best current compression & (global) alignment algorithms give very similar numerical values for MI, when sequences are rather similar:
  → confirms compression method, gives objective and a priori (model-free) success criteria for alignment.