Genes, bioinformatics and dynamics

Peter R Wills

Department of Physics
University of Auckland

Supported by the Alexander von Humboldt Foundation

A different perspective

Metaphysics of science

The Greeks' question:

What remains the same through change?

Post-Einsteinian answer:

Energy/matter as *Urstoff*. Conserved quantity (undifferentiated).

The "gene"

Like an atom. An irreducible element of hereditary. A material particle, but an information carrier.

But what does a gene *do*?

The gene as a causal agent - a programmer.

The Century of the Gene
Evelyn Fox Keller (2000)

1900 De Vries, Correns & Tschermak 'rediscover' Mendel's rules

1906 Bateson coins the term 'genetics'

1909 Johannsen uses the word 'gene' [unit factors, elements, allelomorphs in gametes]

The century of the gene

1943 Avery, McCleod & McCarty identify DNA as carrier of specificity in bacteria

1953 Watson & Crick's DNA structure

2000 Sequence of human genome

Schrödinger, 1944

"How are we to understand that [Hapsburger Lippe] has remained unperturbed by the disordering tendency of the heat motion for centuries?"

The genetic material as an "aperiod crystal", an idea of Delbrück's (1935)

Biological information – the old paradigm

Crick, 1958

Central dogma: DNA → RNA → protein

The Central Dogma ... states that once 'information' has passed into protein *it cannot get out again*. ... Information here means the *precise* determination of sequence, either of the bases in the nucleic acid or of amino acid residues in the protein.

Sequence hypothesis

The specificity of a piece of nucleic acid is expressed solely by the sequence of its bases, and this sequence is a (simple) code for the amino acid sequence of a particular protein.

A fuller picture

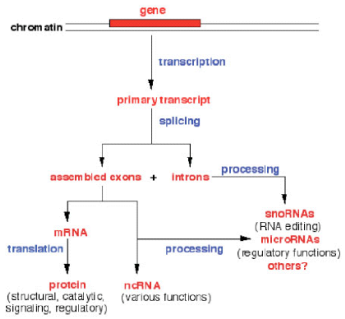
DNA → RNA → protein → function

The real genotype-phenotype mapping occurs in the last (implicit) step of biological information flow.

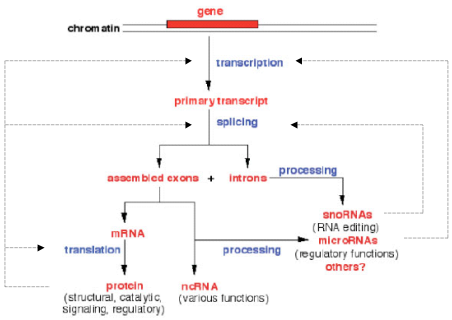
It is governed by:

- folding processes
- structure-function relationship

The new ncRNA paradigm



The new ncRNA paradigm



Bioinformatics

"We have coined the term **Bioinformatics** for the study of informatic processes in biotic systems. Our Bioinformatic approach typically involves spatial, multi-leveled models with many interacting entities whose behaviour is determined by local information." (1978)

Prof. dr. Paulien Hogeweg
Bioinformatics group, Utrecht University

Bioinformatics

The study of the

– generation

– transfer

– coding

of biological information

Importance of dynamics

Informational stability

in biological systems

depends on

dynamic stability.

Importance of dynamics

Informational (in)stability

in biological systems

depends on

(thermo)dynamic (in)stability.

Informational (in)stability

hereditary transfer (mutation/selection)

gene expression (regulation)

translation (maintenance/control)

signalling (cascading)

Replication/autocatalysis

A + components → A + A

Replication/autocatalysis

A + components → A + A

components $\xrightarrow{A}$ A

Replication/autocatalysis

A + components → A + A

components $\xrightarrow{A}$ A

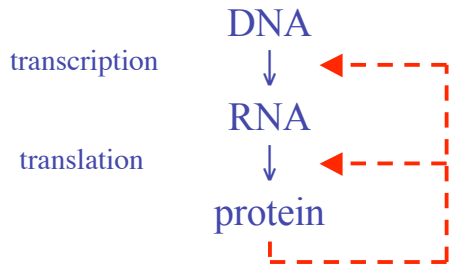
components $\xrightarrow{A}$ A

Replication/autocatalysis

Simplest dynamics for information transfer

nucleic acid replication; quasi-species selection

Autocatalysis in protein synthesis



Autocatalysis in protein synthesis -the information bootstrap problem

If (prebiotic) proteins are produced with more or less random sequences, how can the special "assignment catalysts" needed for a code be selected?

Restatement of the problem

You cannot make protein catalysts with high assignment specificity unless you can construct specific structures (amino acid sequences) accurately. For that you need high assignment specificity.

Why don't errors in translation lead to the catastrophic collapse of the process?

Binary nucleic acid to protein assignments (code).

x = codon; **y** = amino acid

K → k **L → k**

K → l **L → l**

Coding self-organization

In straightforward simulations of translation, the proteins that perform the two coding assignments (**K → k**, **L → l**) are selected and the proteins that perform the twelve "unwanted" assignments (**K → l**, **L → k**) die out.

HOW??

RNA-sequence-dependent protein synthesis

LLKKKKLKK | KLKKKKKKL

↓ ↓

kllkllkll **kllkkklkl**

Which assignments do these proteins catalyse?

"Reflexive" information

We are assuming that proteins are produced from RNA in a collinear fashion (codon frame is 'read" sequentially).

In simulations, the information provided is such that when it is translated by proteins that perform coding assignments, those very same coding-assignment catalysts are produce

– the process is then autocatalytic.

(But how is reflexive information selected?)

Catalytic centres

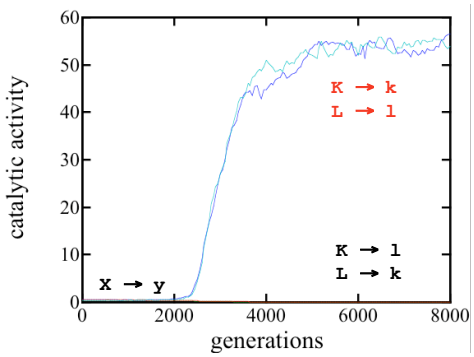
K→k	**llkkkklkkllk**
K→l	**kkkllllkkkll**
L→k	**lkklkkllkkl**
L→l	**klkkkkkllllk**

In a random population of proteins, all assignments (of codons to amino acids) will be catalysed at equal rates and the further random sequences will be synthesised.

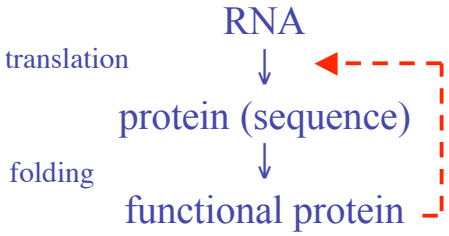
BUT

In the presence of reflexive information, this situation may dynamically unstable, (depending on the way in which catalytic function varies with protein sequence).

Selection of coding catalysts



Coding self-organisation



Problem 1

Where does the reflexive information come from? How can reflexive information be preserved from loss through mutation?

Ans. The coevolution of reflexive information and coding can be stabilised, including protection against parasitism, by co-localisation of nucleic acid information carriers and protein assignment catalysts.

Füchslin & McCaskill, 2002

Problem 2

Qu. Isn't it impossibly improbable that the reflexive information needed for a code as complicated as the one we have (64 codons, 20 amino acids) could be self-selecting because of dynamic instability in a mutually generating population of random nucleic acids and proteins?

Ans. Yes!

What then?

Nucleic acid and protein decomposition

K = {A,B}; L = {B,C}

k = {a,b}; l = {c,d}

X = {A,B,C,D}

y = {a,b,c,d}

For example …

K = GYN, etc. **L** = YRN, etc.

k = hydrophobic **l** = hydrophilic

K = {A, B}; **L** = {C, D}

k = {a, b}; **l** = {c, d}

A = GCN; **B** = other GYN, etc.

a = ala; **b** = other hydrophobic

Binary nucleic acid to protein assignments (code).

X = codon; **y** = amino acid

K → k **L → k**

K → l **L → l**

Code decomposition

A → a B → a **K → k** A → b B → b	C → a D → a **L → k** C → b D → b
B → c B → c **K → l** B → d B → d	C → c D → c **L → l** C → d D → d

Coding assignment set

A → a	B → a	C → a	D → a
A → b	B → b	C → b	D → b
B → c	B → c	C → c	D → c
B → d	B → d	C → d	D → d

Catalytic centres (binary assignments)

K→k	llkkkklkkllk
K→l	kkkllllkkll
L→k	lkklkkllkkl
L→l	klkkkkllllk

Decomposition

K→k	llkkkklkkllk

↓

A→a	ddbbabcaadca
A→b	cdbbbbcabcda
B→a	dcaaaadaaccb
B→b	cdababdbadca

Catalytic centres (quarternary assignments)

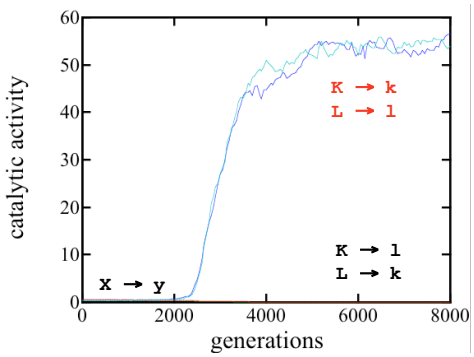
A→a	ddbbabcaadca
A→b	cdbbbbcabcda
B→a	dcaaaadaaccb
B→b	cdababdbadca

A→c	baacddddbadc
A→d	abacdddcbbdc
B→c	bbbcdcccbacc
B→d	baadcddcbbdc

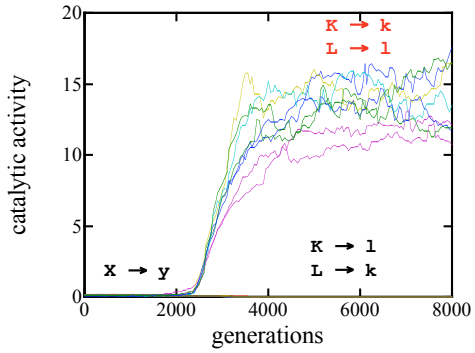
C→a	dbadabcddbbc
C→b	dabdabdcdabd
D→a	daacbbcccaad
D→b	cabdbbdcdbac

C→c	bcbaabbbccdb
C→d	bcababaaddcb
D→c	bdbbabbaccdb
D→d	acabbabaddca

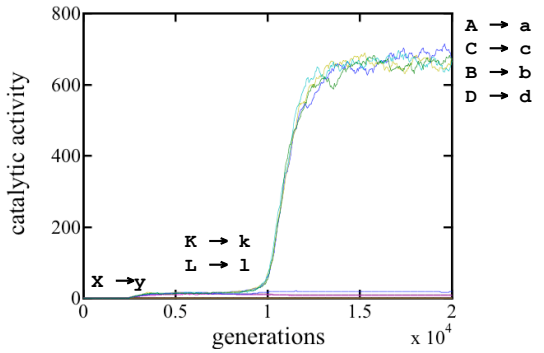
Selection of binary coding catalysts



Selection of binary code catalysts



Transition to quarternary code



Coded RNA-dependent protein synthesis using reflexive information

LLKKKKLKK | KLKKKKKKL

↓ ↓

kllkllkll | kllkkklkl

K → k L → l

Coded RNA-dependent protein synthesis using reflexive information

LLKKKKLKKLLK | KLKKKKKKLLLK

DDBBABCAADCA | CDABABDBADCA | BCBAABBBCCDB | ACABBABADDCA

↓ ↓ ↓ ↓

ddbbabcaadca cdababdbadca bcbaabbbccdb acabbabaddca

A → a B → b C → c D → d

llkkkklkkllk klkkkkkkllllk

K → k L → l

Coded RNA-dependent protein synthesis using reflexive information

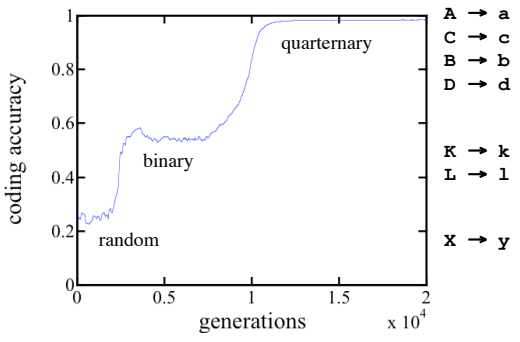
DDBBABCAADCA CDABABDBADCA BCBAABBBCCDB ACABBABADDCA

↓ ↓ ↓ ↓

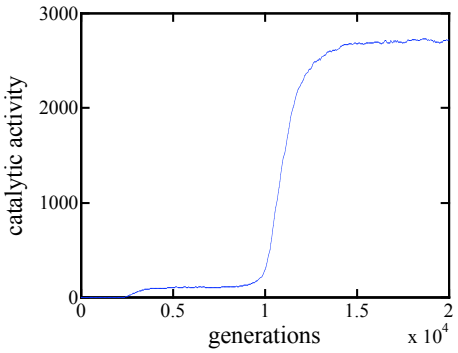
ddbbabcaadca cdababdbadca bcbaabbbccdb acabbabaddca

A → a B → b C → c D → d

Stepwise Evolution of code



Thermodynamic driving force



Coding self-organisation with guided folding

RNA

translation ↓

protein (sequence)

folding ↓

functional protein

Folding process

Protein sequence

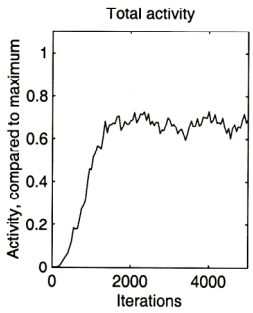
↓

Folded structure

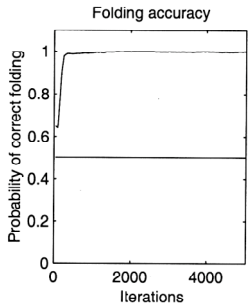
↓

Function (catalytic activity)

Coding self-organisation



Simultaneous folding self-organisation



Susceptibility to parasites

Cooperative autocatalytic systems exhibit susceptibility to parasites that rely on the functionality of the whole system in order to reproduce but which contribute nothing to the operation of the system.

Folding acts as a kind of "scrambling" mechanism that requires parasites to become more and more specialised in order to reproduce (c.f. prions).

Dynamic determination

The biological "meaning" of polymeric sequence information depends on:

– coding
– folding

The operation of these processes depends on the maintenance of a state of self-organization. Thus the outcome of these processes is maintained and determined dynamically. It is not genomically encoded.

Contingency of biological function

Biological function is contingent on

– genetic information
– thermodynamic self-organization

This latter aspect has been generally ignored in bioinformatic studies:

– genetic phylogenies are established on the assumption that there is an unchanging "interpreter" that maps genotype onto phenotype

The real challenge for post-genomic bioinformatics!

To discover the how the molecular biological "interpreter" has evolved.