Languages of Genetic Information

Why do living organisms have 4 nucleotide bases and 20 amino acids?

Apoorva Patel

Centre for Theoretical Studies
and
Supercomputer Education and Research Centre
Indian Institute of Science
Bangalore-560012, India

"The pleasure of finding things out"

The purpose of a language is to communicate information.

It is desirable not to waste physical resources (space, time, energy, . . .) during the process.

How does one find the optimal language for a given task?

A Computer Designer's Point of View:

Software: What is the task?
What is the algorithm?

Hardware: How are the operations implemented?

Efficiency of information processing depends on both software and hardware.

Technical Terms:

Data: They describe a particular realisation of the physical system, amongst its many possible states.

Information: It is the abstract mathematical property obtained by detaching all the physical characteristics from data.

Knowledge: It is obtained by adding a sense of purpose to the abstract information.

Information = Data - Physical Realisation

Knowledge = Information + Interpretation

1. Abstract information can be manipulated with precise mathematical rules, without going into nitty-gritty of its meaning.
2. The manipulations can only be implemented using physical devices.
3. The interpretation of a language has to be established through physical properties.

Abstract information theory does not tell us what physical realisation would be appropriate for a particular message, nor does it tell us the best way of implementing a computational task. These choices have to be made by analysing the type (and not amount) of information, and inspecting the available physical resources.

Biological Facts:

- Languages of genes and proteins are universal:
  The same 4 nucleotide bases and 20 amino acids are used in DNA, RNA and proteins, all the way from viruses and bacteria to human beings. This is despite the fact that other nucleotide bases and amino acids exist in the cell.
  $\Rightarrow$ Selection of a specific language has taken place.
- Genetic information is encoded close to data compression limit and maximal packing.
  $\Rightarrow$ Optimisation of information storage has taken place.
- Evolution occurs through random mutations, which are local changes in the genetic sequence. Only a small fraction of the mutations survive.
  Darwinian selection is the optimising mechanism.

Local changes can get trapped in locally optimal configurations, without reaching the global optimum. History is crucial in the process, and the process typically produces many surviving candidates.

The globally optimal solution is easier to reach, when the number of possibilities is small and/or the range of exploration is large.

Frozen accident? No!

The language somehow came in to existence, and became such a vital part of life's machinery that any change in it would be highly deleterious to living organisms.

Requires an extremely rare event, without sufficient time to explore other possibilities.

Optimal solution? Yes!!

The language arrived at its best form by trial and error, and it did not change thereafter, because any change in it would make the communication process worse.

Requires competition amongst many possibilities, where the optimal solution wins over all other options.

After the discovery of DNA structure, many people tried to construct an "optimal solution" scenario, and failed. Biologists then by and large accepted the "frozen accident" scenario.

What has caused the recent rethinking?

1) Large amount of information has become available in gene and protein databases, which can be used to test hypotheses.

2) Explosive growth of computer science and technology has led to a better understanding of optimisation criteria in information processing.

Evolution:

Organism	Messages	Physical Means
Single cell	Molecular (DNA, Proteins)	Chemical bonds, Diffusion
Multicellular	Electrochemical (Nervous system)	Convection, Conduction
Families, Societies	Imitation, Teaching, Languages	Light, Sound
Humans	Books, Computers, Telecommunication	Storage devices, Electromagnetic waves
Cyborgs ?	Databases	Merger of brain computer

Evolution has progressively discovered higher level of communication mechanisms.

- Communication range expands.
- Physical contact reduces.
- Abstraction increases.
- Succinct language forms arise.
- Complex translation machinery develops.

"Knowledge" is the driving force behind "survival of the fittest".
It provides direction for evolution.

Hierarchical Processing of Information:

Computers		Living organisms
Data	Input	Environmental signals
Pre-processor	High level	Sense organs
Compiler	⇑	Nervous system
Assembler	Translation	Brain
	⇓	
Machine code	Low level	Electrochemical signals
Electrical signals	Execution	Proteins
Programmer	Programme	DNA

High level processing is abstract, with a wide variety of instructions and subjective adaptations.

Low level processing is directly related to physical properties, with a limited number of instructions and tasks.

At the lowest level of information processing, the physical objects that carry the message have to convey the information as well as its interpretation.

The Tasks:

DNA/RNA: Assemble a chain of building blocks on top of a pre-existing master template.
(DNA is the read-only-memory of living organisms.)

Proteins: Design structurally stable molecules of specific shapes, with precise locations of active chemical groups.
(Proteins carry out various functions, by highly specific binding—lock-and-key mechanism.)

Optimisation Criteria:

- Minimise errors: Use clearly distinguishable building blocks, with discrete operations. Practical applications need only bounded error calculations.
- Minimise resources: Use a small number of building blocks, with simple and quick operations. In a versatile language, the building blocks can be joined together in as many different ways as possible.

The language with the smallest set of building blocks (for a given task) has a unique status in the optimisation procedure:

- Largest tolerance against errors.
- Smallest instruction set.
- High density of packing.
- Simplest language, without need of translation.

Information Systems and Dimensionality:

Generalise the concept of language, from a "sequence of letters" to a "collection of building blocks".

For a d-dim system, the simplest building block is a simplex, i.e. a set of $(d+1)$ points.

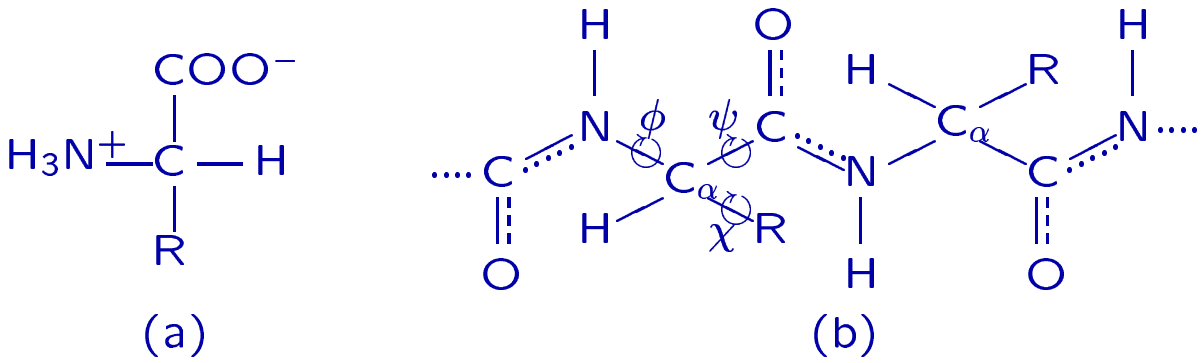
Group representations and transformations fix the structure of the language.

	1-dim	3-dim
Information	Numerical	Structural
Discrete space	Integers	Lattice
Basic variables	$\{0,1\}$	$l = \{0,1\}$
Implementation	On/Off	Tetrahedron
Operations	Addition, Multiplication	Translation, Rotation

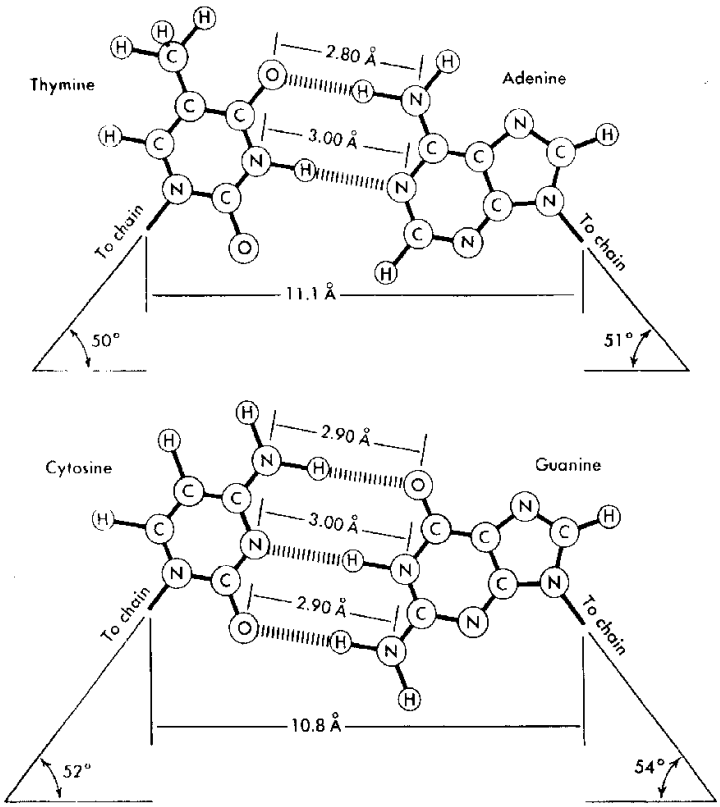
Folded Chains:

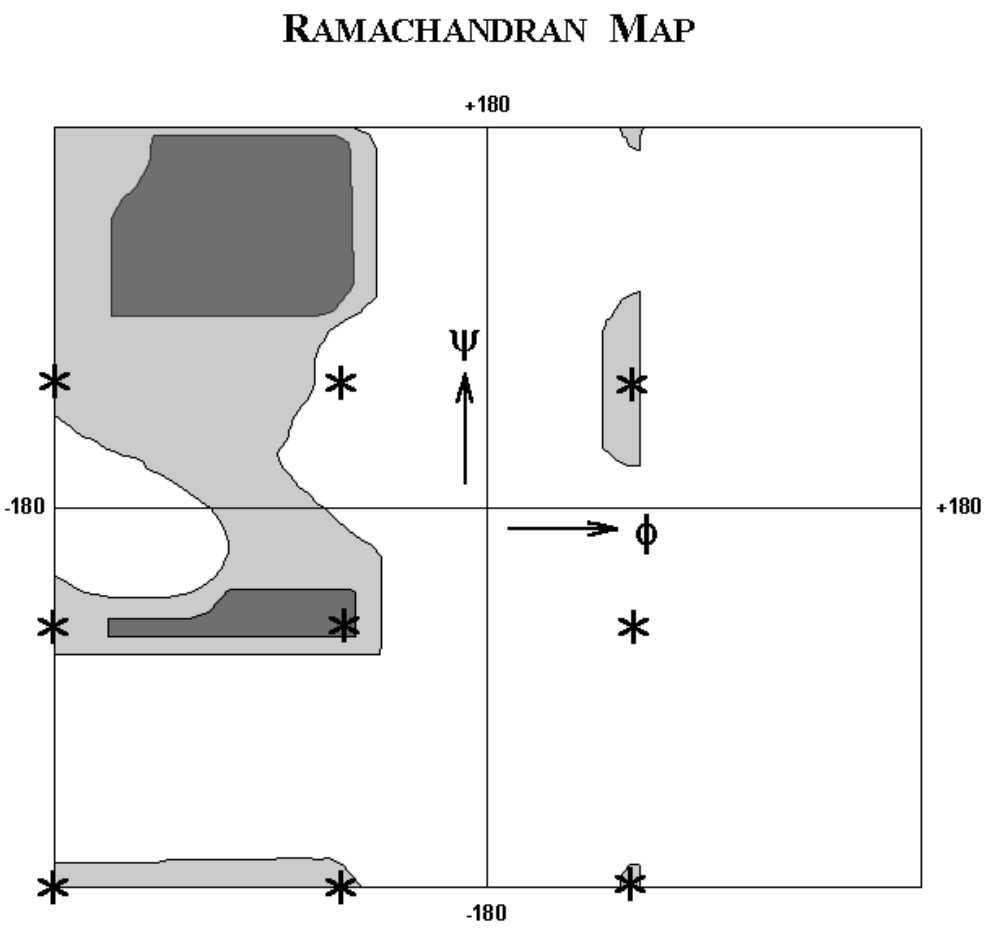
- 1-dim sequences are easier to synthesise than 3-dim structures.
- The chain itself can carry information about bends and folds.
- A variety of folding options are available for a given structure.
- Folded $\Leftrightarrow$ unfolded transition requires flexible joints and weak non-local interactions (close to critical behaviour).

Polypeptide Chains:



Nucleotide Base-pairings:





The Ramachandran map for chiral L-type amino acids, displaying the allowed orientation angles for the C_α bonds in real polypeptide chains, after taking in to account hard core repulsion between all the atoms (Ramachandran *et al.* 1963). The angles ϕ and ψ are periodic. In the approximation that embeds the polypeptide chain on a diamond lattice, only nine discrete possibilities exist for the angles. These are marked by stars on the same plot; they are uniformly separated by 120° steps.

Amino acid	R-group property	Mol. wt.	Class
Gly (Glycine)	Non-polar	75	II
Ala (Alanine)	Non-polar	89	II
Pro (Proline)	Non-polar	115	II
Val (Valine)	Non-polar	117	I
Leu (Leucine)	Non-polar	131	I
Ile (Isoleucine)	Non-polar	131	I
Ser (Serine)	Polar	105	II
Thr (Threonine)	Polar	119	II
Asn (Asparagine)	Polar	132	II
Cys (Cysteine)	Polar	121	I
Met (Methionine)	Polar	149	I
Gln (Glutamine)	Polar	146	I
Asp (Aspartate)	$-$ve charge	133	II
Glu (Glutamate)	$-$ve charge	147	I
Lys (Lysine)	$+$ve charge	146	II
Arg (Arginine)	$+$ve charge	174	I
His (Histidine)	Ring/Aromatic	155	II
Phe (Phenylalanine)	Ring/Aromatic	165	II
Tyr (Tyrosine)	Ring/Aromatic	181	I
Trp (Tryptophan)	Ring/Aromatic	204	I

Properties of the amino acids depend on their side chain R-groups. Larger molecular weights indicate longer side chains. The 20 amino acids naturally occurring in proteins have been divided into two classes of 10 each, depending on the properties of aminoacyl-tRNA synthetases that bind the amino acids to tRNA. These classes divide amino acids with each R-group property equally, the longer side chains correspond to class I and the shorter ones correspond to class II.

Summary (Proteins):

1. What is the purpose of the language of amino acids?
   To form protein molecules of different shapes and sizes, and containing different chemical groups.
2. What is the best discrete geometry for designing 3-dim structures?
   Tetrahedral geometry and diamond lattice.
   (Secondary protein structures—α-helices, β-bends, β-sheets—fit easily on the diamond lattice.)
3. What are the ideal physical components to realise this geometry?
   Covalently bonded carbon atoms.
   (Also N^+ and H_2O.)
4. What is a convenient way to assemble these components in the desired structure?
   Synthesise 1-dim polypeptide chains, which contain information about how to fold in to 3-dim structures.
5. What are the elementary operations needed to fold a polypeptide chain on a diamond lattice, in any desired manner?
   Nine discrete rotations, represented as 3×3 array on the Ramachandran map. Trans-cis flip and long distance bonds are additional operations.
6. What can the side-chain R-groups do?
   They favour particular orientations by interactions amongst themselves. They fill up cavities in the structure by variations in their size.

What are the properties of amino acids?

(a) The 20 amino acids are divided in to two classes of 10 each, by the bilingual aminoacyl-tRNA synthetases.

(b) The duplication is a binary label for the size of the side chain R-group.

(c) Each class contains a special amino acid.

- Cys in class I makes long distance disulfide bonds.
- Pro in class II can induce trans-cis flip.

Possible Evolution of the Genetic Code:

Continuity has to be maintained during evolution. Drastic changes kill living organisms.

The present non-overlapping triplet genetic code arose from a more primitive doublet one.

(1) In the 10 amino acid doublet genetic code, the third base was a non-coding separation mark. The cavities of protein structure were not filled completely.

(2) The third base was put to use as a double-valued classical label for the R-group size. This increased structural stability of proteins by filling cavities.

(3) Further optimisation juggled codons, and introduced START/STOP signals. Similar codons for similar amino acids, and the wobble rules, are relics of the doubling of the genetic code, indicative but not perfect.

A two party game: "Name the person"

One team chooses the name of a well-known person.

The other team has to discover the name by asking questions to the first team.

The first team provides only binary answers to the questions (YES or NO), i.e. it releases minimal information.

The two teams take turns choosing persons and asking questions.

After several rounds, the team that discovers the names using less number of questions is the winner.

Strategy:

Highly specific Questions such as "Is the person Albert Einstein?" are inefficient. They fail most of the time.

Efficient questions are of the type "Is the person a man or a woman?", which reduce the number of possibilities by two at every step.

The computer science paradigm for this game is "Database search", and binary search is the optimal classical algorithm.

A sorted database of N items can be searched using $\log_2 N$ binary questions.

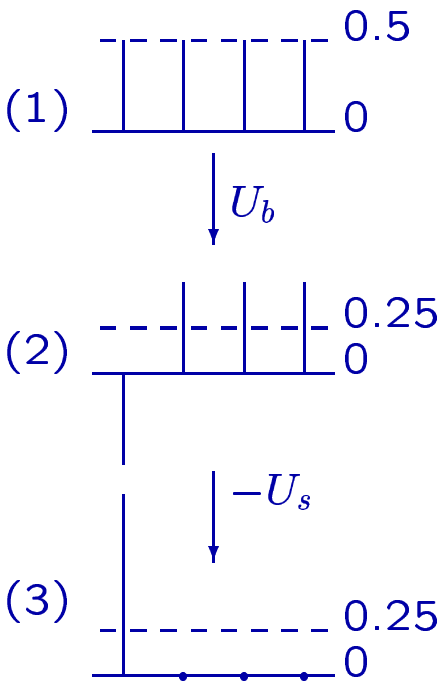
An unsorted database of N items can be searched using $N/2$ binary questions with memory, and using N binary questions without memory.

Biochemical Assembly Process:

- Instead of waiting for a desired complex biomolecule to come along, it is far more efficient to synthesise it from common, simple ingredients.
- There should be a sufficient number of clearly distinguishable building blocks to create the wide variety of required biomolecules.
- The database of building blocks is unsorted. The assembly of DNA, RNA and polypeptide chains takes place on pre-existing templates.
- At each step in the linear assembly process, base-pairings decide (by complementarity rule) which building block is the correct one to add.
- Molecular bonds are binary questions; either they form or do not form.
- Classically two nucleotide bases (one complementary pair) are sufficient to encode genetic information. That would have in any case preceded (during evolution) the four nucleotide base system found in nature.
- Optimal language would minimise resources for the assembly process, in space as well as in time, and would be favoured by Darwinian selection.
- Laws of quantum mechanics are unavoidable in molecular processes, and their optimal solutions are different than the classical ones.

Quantum Database Search:

The steps of the algorithm for the simplest case of 4 items in the database, when the first item is desired by the oracle.



Uniform distribution	Equilibrium configuration
Quantum oracle	Binary question
Amplitude of desired state flipped in sign	Sudden perturbation
Reflection about average	Overrelaxation
Desired state reached	Opposite end of oscillation
Algorithm is stopped	Measurement

The left column depicts the amplitudes of the 4 states, with the dashed lines showing their average values. The middle column describes the algorithmic steps, and the right column mentions their physical implementation.

Summary (DNA):

1. What is the information processing task carried out by the genetic code?
   Assembling molecules by picking up components from an unsorted database.
2. What is the optimal way of carrying out this task?
   Lov Grover's quantum search algorithm.
   (Requires wave mechanics.)
3. What is the signature of this algorithm?

$$(2Q+1)\sin^{-1}\frac{1}{\sqrt{N}} = \frac{\pi}{2} \implies \begin{cases} Q=1, & \text{N=4} \\ Q=2, & \text{N=10.5} \\ Q=3, & \text{N=20.2} \end{cases}$$

4. Does the genetic machinery have the structure to implement this algorithm?
   Yes!
   All this means that if we have to design a system to implement this task, knowing all the physical laws that we do, we would opt for something like what is present in nature.
5. What did nature do? Was this algorithm exploited when the genetic code evolved billions of years ago?
   "?" (Evolution of life cannot be repeated.)
6. Do living organisms use this algorithm even today?
   In principle, experimentally testable!

Decoherence and stability of the algorithm:

Fermi's Golden Rule:
Cross – section = |Matrix Element|^2·Flux·Density of states

- Though Grover's algorithm was discovered in the context of quantum computation, it can be implemented using a set of wave modes.
  At the molecular scale, in addition to quantum dynamics, there exist wave modes corresponding to rotations and vibrations.
- The number of oracle calls remains fixed, but physical resources required to implement the algorithm depend on the manner of implementation.
  Quantum: $\log_2 N$ q-bits, entangled with each other
  Waves: N modes without entanglement
- In presence of environmental disturbances, entanglement is far more fragile than mere superposition.
- No wave motion can be damped faster than its natural undamped oscillation frequency. Too much damping freezes the motion (Zeno effect).

$$\ddot{x} + 2\gamma\dot{x} + \omega_0^2 x = 0, \;\; x \sim e^{i\omega t}$$
$$\Rightarrow \;\; \gamma_{\text{crit}} = \max(\text{Im}\omega) = \omega_0$$

- For atomic phenomena, $\omega_0 = \Delta E/\hbar = O(10^{14})\text{sec}^{-1}$.

References:

All my papers are available at http://arXiv.org/

quant-ph/0002037
quant-ph/0102034
quant-ph/0105001
quant-ph/0103017
quant-ph/0202022 (review)
quant-ph/0206014
quant-ph/0306158

I am also writing a book on the subject, to be published by World Scientific, Singapore.