Towards Understanding the Origin of Genetic Languages

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

Apoorva Patel

Centre for High Energy Physics and

Supercomputer Education and Research Centre

Indian Institute of Science, Bangalore

26 January 2013, IIT, Jodhpur

What is Life?

Life is fundamentally a non-equilibrium process, commonly charaterised in terms of two basic phenomena.

What is Life?

Life is fundamentally a non-equilibrium process, commonly charaterised in terms of two basic phenomena.

Metabolism: Many biochemical processes are needed to sustain a living organism. That requires a continuous supply of free energy, extracted from the environment. (The ultimate source of this energy is gravity—the only interaction that is not in equilibrium.)

Reproduction: A particular structure cannot survive forever, because of environmental disturbances and damages. So life perpetuates itself by a succession of generations.

What is Life?

Life is fundamentally a non-equilibrium process, commonly charaterised in terms of two basic phenomena.

Metabolism: Many biochemical processes are needed to sustain a living organism. That requires a continuous supply of free energy, extracted from the environment. (The ultimate source of this energy is gravity—the only interaction that is not in equilibrium.)

Reproduction: A particular structure cannot survive forever, because of environmental disturbances and damages. So life perpetuates itself by a succession of generations.

Both these phenomena are sustaining/protecting/improving something, and that too against the odds.
But what is it that is being sustained/protected/improved?

The meaning of it all

All living organisms are made up of atoms.
Atoms are fantastically indestructible.
They just get rearranged in different ways.
Each of us would have a billion atoms that once belonged to the Buddha, or Genghis Khan, or Isaac Newton.

The meaning of it all

All living organisms are made up of atoms.
Atoms are fantastically indestructible.
They just get rearranged in different ways.
Each of us would have a billion atoms that once belonged to the Buddha, or Genghis Khan, or Isaac Newton.

It is not the atoms themselves but their arrangements, which carries biochemical information.
Living organisms evolve, by altering atomic arrangements.
Molecules keep on changing, and atoms are shuffled.

The meaning of it all

All living organisms are made up of atoms.
Atoms are fantastically indestructible.
They just get rearranged in different ways.
Each of us would have a billion atoms that once belonged to the Buddha, or Genghis Khan, or Isaac Newton.

It is not the atoms themselves but their arrangements, which carries biochemical information.
Living organisms evolve, by altering atomic arrangements.
Molecules keep on changing, and atoms are shuffled.

Hardware is recycled,
while software is improved!

Preservation of information requires complex structures!!

Typical scales:

Atoms: H, C, N, O, and infrequently P, S.

Nucleotide bases and amino acids: 10-20 atoms

Peptides and drugs: 40-100 atoms

Proteins: 100-1000 amino acids

Genomes: 10^3-10^9 nucleotide base pairs

Size: 1 nm (molecules)-10^4 nm (cells)

Typical scales:

Atoms: H, C, N, O, and infrequently P, S.

Nucleotide bases and amino acids: 10-20 atoms

Peptides and drugs: 40-100 atoms

Proteins: 100-1000 amino acids

Genomes: 10^3-10^9 nucleotide base pairs

Size: 1 nm (molecules)-10^4 nm (cells)

Gene and protein databases are accumulating a lot of information, which can be used to test hypotheses and consequences of optimised information processing.

We hope to understand physical evolutionary reasons for (1) specific languages, and (2) their specific realisations.

These have a bearing on probability of finding life elsewhere in the universe.

Biological Facts

1. 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 living cells.
   $\Rightarrow$ Selection of a specific language has taken place.

Biological Facts

1. 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 living cells.
   $\Rightarrow$ Selection of a specific language has taken place.

2. Genetic information is encoded close to
   the data compression limit (for genes)
   and maximal packing (for proteins).
   $\Rightarrow$ Optimisation of information storage has occurred.

Biological Facts

1. 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 living cells.
   $\Rightarrow$ Selection of a specific language has taken place.
2. Genetic information is encoded close to
   the data compression limit (for genes)
   and maximal packing (for proteins).
   $\Rightarrow$ Optimisation of information storage has occurred.
3. Evolution occurs through random mutations,
   which are local changes in the genetic sequence.
   Only a small fraction of the mutations survive.
   Darwinian selection (competition for finite resources
   among the users) is the optimising mechanism.

Two Scenarios

Frozen accident? No!
The language somehow came into 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 information processing worse.
Requires sufficient time to generate many possibilities,
and subsequent competition amongst them.
The optimal solution then wins over all other options.

The Tasks

DNA/RNA: Assemble a chain of building blocks on top of a pre-existing master template.
DNA is the read-only-memory (ROM) of living organisms.
This is a 1-dimensional problem.

The Tasks

DNA/RNA: Assemble a chain of building blocks on top of a pre-existing master template.
DNA is the read-only-memory (ROM) of living organisms.
This is a 1-dimensional problem.

Proteins: Design structurally stable molecules of specific shapes, with precise locations of active chemical groups.
Proteins carry out various functions, by highly specific binding and short range interactions (hard-core repulsion, screened charges, van der Waals forces, hydrogen bonds), i.e. lock-and-key mechanism.
The key shape accuracy is a fraction of atomic size.
This is a 3-dimensional problem.

Optimisation Criteria

Minimise errors: Use a digital language—clearly distinguishable building blocks, with discrete operations.
Discrete values are associated with islands of physical properties, instead of precise points.
Practical applications need only bounded error calculations.

Optimisation Criteria

Minimise errors: Use a digital language—clearly distinguishable building blocks, with discrete operations.
Discrete values are associated with islands of physical properties, instead of precise points.
Practical applications need only bounded error calculations.

Minimise resources: Use a small number of building blocks, which allow simple and quick operations.
In a versatile language, the building blocks can be joined together in as many different ways as possible, giving rise to distinct structures.

Top-down vs. Bottom-up

Michelangelo's David

Top-down vs. Bottom-up

Michelangelo's David

Lego Construction Blocks

Minimal Language

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.

(Discrete variables are spread as far apart as possible in the available range of physical hardware properties.)

- Smallest instruction set.

(Number of possible transformations is limited.)

- High density of packing and quick operations.

(These more than make up for the increased depth of computation.)

- Simplest language, without need of translation.

(Simple physical responses of the hardware can be used.)

Boolean Algebra

It is the minimal classical language for encoding information as 1-dimensional sequences.

Its two letters can have a variety of realisations:
0 and 1, on and off, up and down, etc.

Its operations form the smallest field Z_2.

It is sufficient (though not necessarily optimal) to encode any information, as our computers demonstrate.

When starting from scratch (e.g. at the origin of life), it is the simplest language to arrive at.

Boolean Algebra

It is the minimal classical language for encoding information as 1-dimensional sequences.

Its two letters can have a variety of realisations:
0 and 1, on and off, up and down, etc.

Its operations form the smallest field Z_2.

It is sufficient (though not necessarily optimal) to encode any information, as our computers demonstrate.

When starting from scratch (e.g. at the origin of life), it is the simplest language to arrive at.

So why did evolution opt for more complex languages, and how? We have to understand the optimisation of languages, while implementing specific tasks!

Summary (Proteins)

1. What is the purpose of the language of amino acids?

. . . contd.

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.

. . . contd.

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 optimal discrete geometry for designing 3-dimensional structures (translations and rotations)?

. . . contd.

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 optimal discrete geometry for designing 3-dimensional structures (translations and rotations)?
   Simplicial tetrahedral geometry and diamond lattice.
   (Secondary protein structures, i.e. α-helices, β-bends and β-sheets, fit easily on the diamond lattice.)

   Diamond is the hardest material, with the largest band gap.

. . . contd.

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 optimal discrete geometry for designing 3-dimensional structures (translations and rotations)?
   Simplicial tetrahedral geometry and diamond lattice.
   (Secondary protein structures, i.e. α-helices, β-bends and β-sheets, fit easily on the diamond lattice.)

   Diamond is the hardest material, with the largest band gap.
3. What are the best physical components to realise this geometry?

. . . contd.

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 optimal discrete geometry for designing 3-dimensional structures (translations and rotations)?
   Simplicial tetrahedral geometry and diamond lattice.
   (Secondary protein structures, i.e. α-helices, β-bends and β-sheets, fit easily on the diamond lattice.)

   Diamond is the hardest material, with the largest band gap.

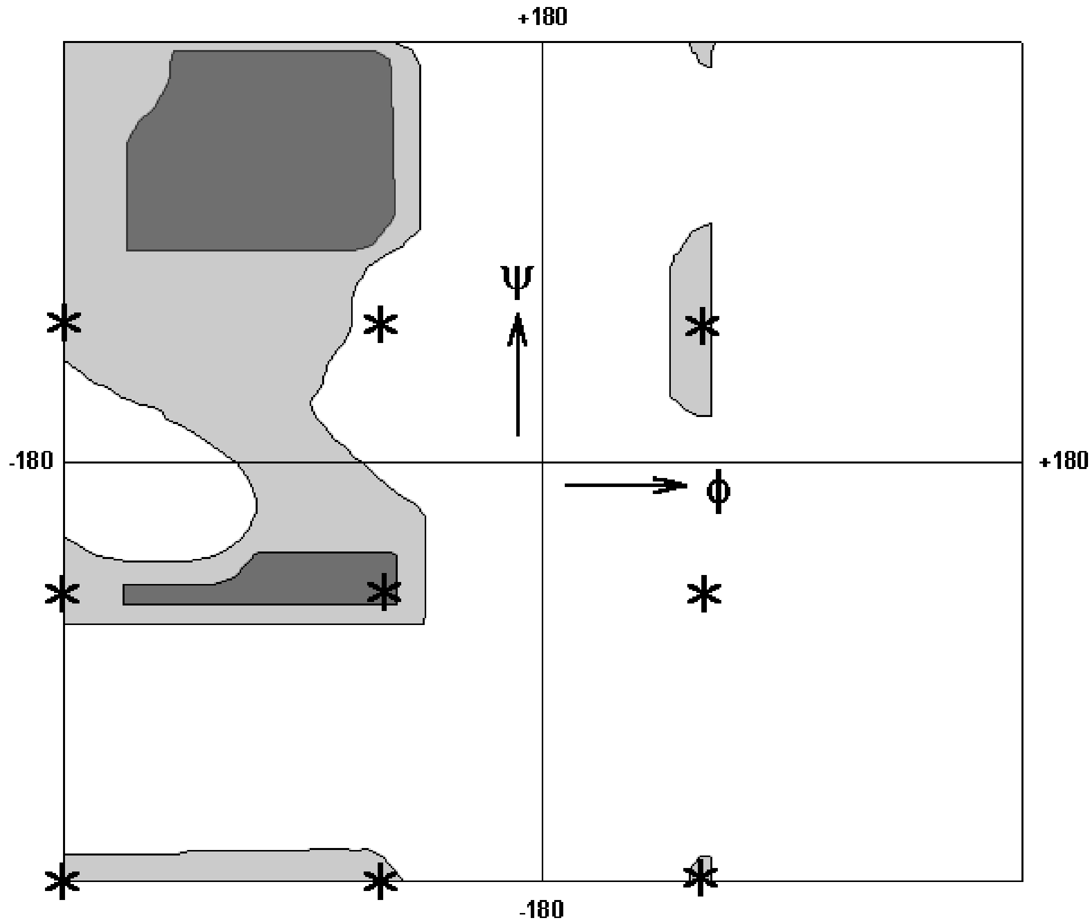
3. What are the best physical components to realise this geometry?
   Covalently bonded carbon atoms. Also N^+ and H_2O.
   (In the graphite sheet form, carbon also provides the simplicial geometry for 2-dim membrane patterns.)

. . . contd.

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 into 3-dim structures.
   The same structure may be covered by different folding patterns,
   so all the folds may not occur with equal probability.

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 into 3-dim structures.
   The same structure may be covered by different folding patterns, so all the folds may not occur with equal probability.
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.

Ramachandran Map



The allowed orientation angles for the C_α bonds in real polypeptide chains for chiral L-type amino acids, taking into account hard core repulsion between atoms. Stars mark the nine discrete possibilities for the angles, when the polypeptide chain is folded on a diamond lattice.

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 into 3-dim structures.
   The same structure may be covered by different folding patterns, so all the folds may not occur with equal probability.

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.
   Chemical properties decide orientation. Physical volume adjusts to cavity size.

Amino acid	R-group property	Mol. wt.	Class	Propensity
G Gly (Glycine)	Non-polar	75	II	turn
A Ala (Alanine)	aliphatic	89	II	α
P Pro (Proline)		115	II	turn
V Val (Valine)		117	I	β
L Leu (Leucine)		131	I	α
I Ile (Isoleucine)		131	I	β
S Ser (Serine)	Polar	105	II	turn
T Thr (Threonine)	uncharged	119	II	β
N Asn (Asparagine)		132	II	turn
C Cys (Cysteine)		121	I	β
M Met (Methionine)		149	I	α
Q Gln (Glutamine)		146	I	α
D Asp (Aspartate)	Negative	133	II	turn
E Glu (Glutamate)	charge	147	I	α
K Lys (Lysine)	Positive	146	II	α
R Arg (Arginine)	charge	174	I	α
H His (Histidine)	Ring/	155	II	α
F Phe (Phenylalanine)	aromatic	165	II	β
Y Tyr (Tyrosine)		181	I	β
W Trp (Tryptophan)		204	I	β

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.

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 dynamics.)

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.
- The assembly process is linear. At each step, base-pairings decide (by complementarity rule) which building block is the correct one to add.
- Molecular bonds are binary questions; either they form or they do not form.

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 dynamics.)
3. What is the signature of this algorithm?

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

...contd.

Database Search

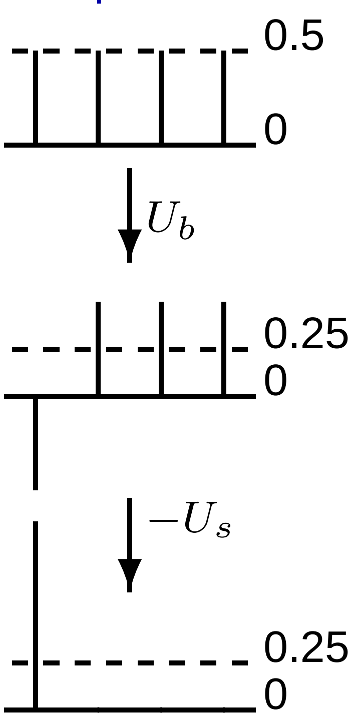
The computer science paradigm for the "Name the person" game is "Database search".

Classical:
Binary tree 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.

Quantum:
Wave mechanics works with amplitudes and not with probabilities. Superposition of amplitudes can yield constructive as well as destructive interference.
Optimal search solutions differ from the classical ones.

Quantum Database Search

The steps of the algorithm for the simplest case of 4 items in the database. Let the first item be desired by the oracle.



	Amplitudes	Algorithmic Steps	Physical Implementation
(1)		Uniform distribution	Equilibrium configuration
	U_b	Quantum oracle	Binary question
(2)		Amplitude of desired state flipped in sign	Sudden perturbation
	$-U_s$	Reflection about average	Overrelaxation
(3)		Desired state reached	Opposite end of oscillation
(4)	Projection	Algorithm is stopped	Measurement

(Dashed line denotes the average amplitude.)

4. Does the genetic machinery have the ingredients to implement this algorithm?
   Yes!
   Superposition can be quantum (e.g. wavefunction), classical (e.g. vibrations), or illusory (selection time longer than transit time between possibilities).

All this means that if we have to design a language to implement this task, knowing all the physical laws that we do, we would opt for something like what is present in nature.

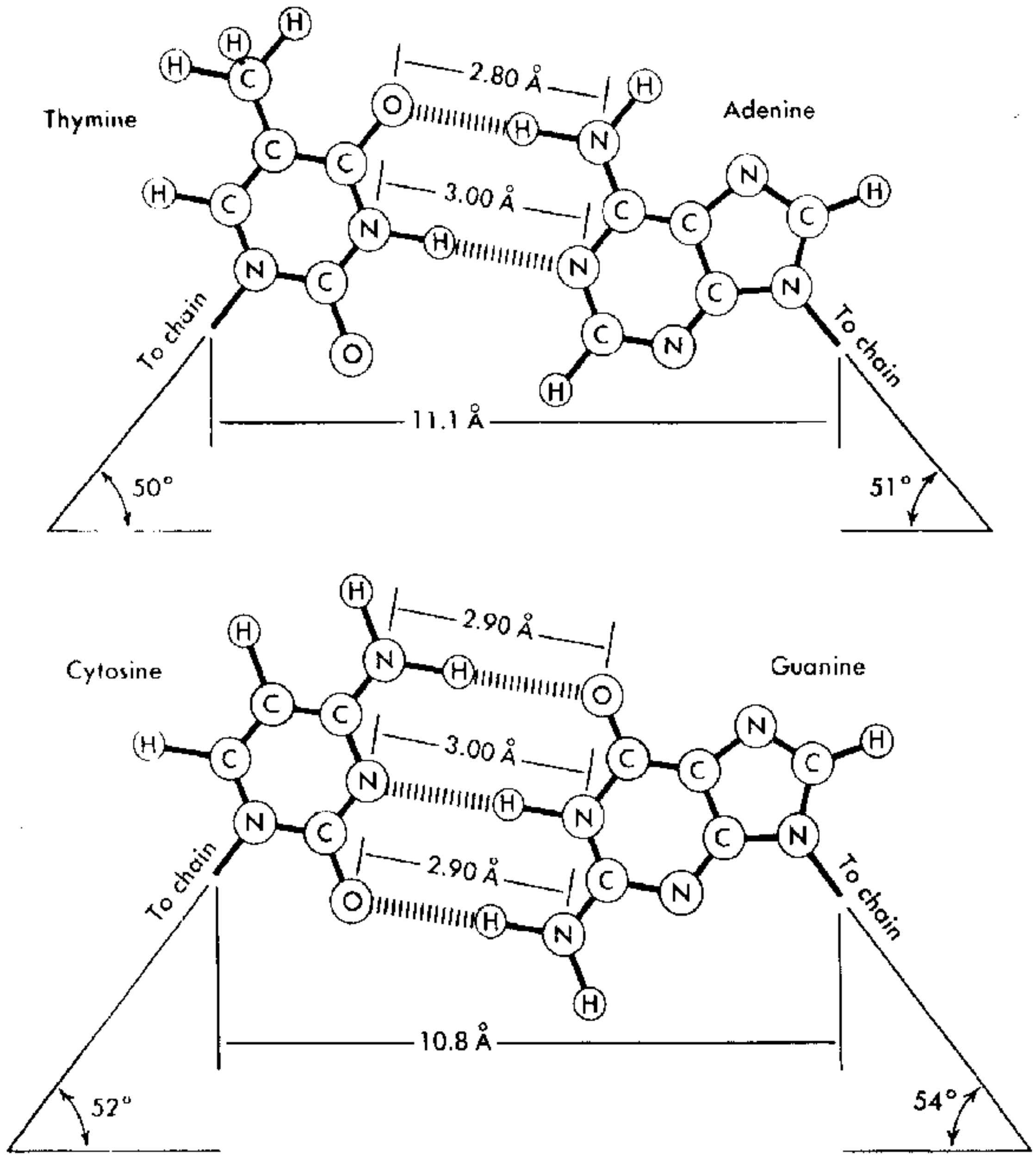
4. Does the genetic machinery have the ingredients to implement this algorithm?
   Yes!
   Superposition can be quantum (e.g. wavefunction), classical (e.g. vibrations), or illusory (selection time longer than transit time between possibilities).

All this means that if we have to design a language to implement this task, knowing all the physical laws that we do, we would opt for something like what is present in nature.

Classically two nucleotide bases (one complementary pair) are sufficient to encode the genetic information.
Such a simpler system should have preceded (during evolution) the four nucleotide base system found in nature.

Was the speed-up provided by the wave algorithm the real incentive for nature to complicate the language?

Nucleotide Base-pairings



4. Does the genetic machinery have the structure to implement this algorithm?
   Yes!
5. What did nature do? Was this algorithm exploited when the genetic code evolved billions of years ago?
   “?”
   (Evolution of life cannot be repeated, and there is a limit to extrapolating evolutionary trees back in time.)

4. Does the genetic machinery have the structure to implement this algorithm?
   Yes!

5. What did nature do? Was this algorithm exploited when the genetic code evolved billions of years ago?
   "?"
   (Evolution of life cannot be repeated, and there is a limit to extrapolating evolutionary trees back in time.)

6. Do living organisms use this algorithm even today?
   In principle, experimentally testable!
   Rely on Darwinian selection: Construct artificial languages having a subset of the building blocks, and make them compete against the natural language.

4. Does the genetic machinery have the structure to implement this algorithm?
   Yes!
5. What did nature do? Was this algorithm exploited when the genetic code evolved billions of years ago?
   “?”
   (Evolution of life cannot be repeated, and there is a limit to extrapolating evolutionary trees back in time.)
6. Do living organisms use this algorithm even today?
   In principle, experimentally testable!
   Rely on Darwinian selection: Construct artificial languages having a subset of the building blocks, and make them compete against the natural language.

Need a believable, and testable, atomic scale model implementing Grover’s algorithm using nucleotide bases.

References

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

quant-ph/0002037:Pramana 56 (2001) 367-381

quant-ph/0102034:J. Genet. 80 (2001) 39-43

quant-ph/0105001:J. Biosc. 26 (2001) 145-151

quant-ph/0103017:J. Biosc. 27 (2002) 207-218

quant-ph/0206014:Fluct. Noise Lett. 2 (2002) L279-284

quant-ph/0202022 (review):Computing and Information Sciences: Recent Trends, Ed. J.C. Misra, Narosa (2003) 271-294

q-bio.GN/0403036:J. Theor. Biol. 233 (2005) 527-532

quant-ph/0503068:Proceedings of QICC 2005, IIT Kharagpur, February 2005, Ed. S.P. Pal and S. Kumar, Allied Publishers (2006) 197-206

quant-ph/0401154:Int. J. Quant. Inform. 4 (2006) 815-825

quant-ph/0609042:AIP Conference Proceedings 864 (2006) 261-272

0705.3895[q-bio.GN] (review):Chapter 10, Quantum Aspects of Life, Eds. D. Abbott et al., Imperial College Press (2008) 187-219