SCALE INVARIANCE EMBEDDED VOTES and SELF-EMERGING BINARY LOGICS in the WHOLE HUMAN GENOME

Revisiting "HIERARCHICAL INTROSPECTIVE LOGICS" John F. NASH's theories Highlighting by the Human Genome self-emerging embedded BINARY LOGICS discovery

By Dr Jean-claude Perez mailto:jeanclaudeperez2@free.fr Bordeaux 17 February 2008,

"There can exist no procedure for finding the set of all regularities of an entity. But classes of regularities can be identified. Finding regularities typically refers to taking the available data about the entity, processing it in some manner into, say, a bit string, ***and then dividing that string into parts in a particular way*** *and looking for mutual AIC (Algorithmic Information Content) among the parts. If a string is divided into two parts, for example, the mutual AIC can be taken to be the sum of the AIC's of the parts minus the AIC of the whole.* ***An amount of mutual algorithmic information content above a certain threshold can be considered diagnostic of a regularity****. Given the identified regularities, the corresponding effective complexity is the AIC of a description of those regularities".*

In **"What Is Complexity?"**

By Pr Murray Gell-Mann ,In John Wiley and Sons, Inc.: *Complexity*, Vol. 1, no.1 © 1995.

http://web.archive.org/web/20011121181321/www.santafe.edu/sfi/People/mgm/complexity.html

I-ABSTRACT :

In one hand, in his unformal paper entitled "HIERARCHICAL INTROSPECTIVE LOGICS" (visit http://www.math.princeton.edu/jfnj/texts_and_graphics/LOGIC/talk.CMU/HIL39e.htm), the Nobel prize Professor John F. Nash Jr. explores new approaches of Turing/Godel undecidability problems adding particularly the **"EMBEDDABILITY" dimension**.
In other hand, we have discovered mathematical CODES structuring all genomes and particularly the whole sequenced FINALIZED HUMAN GENOME (three billions base-pairs about distributed in the 24 Human chromosomes). This discovery must be published, meanwhile the attached abstract describes the evidence of a self-emerging embedded BINARY CODE structuring the whole human genome.
Then, in the second part of the paper, we demonstrate generalization of this emergent binary code to long genomes and particularly to the whole human genome.
Finally, we show that both binary states values have strong links with "golden ratio".

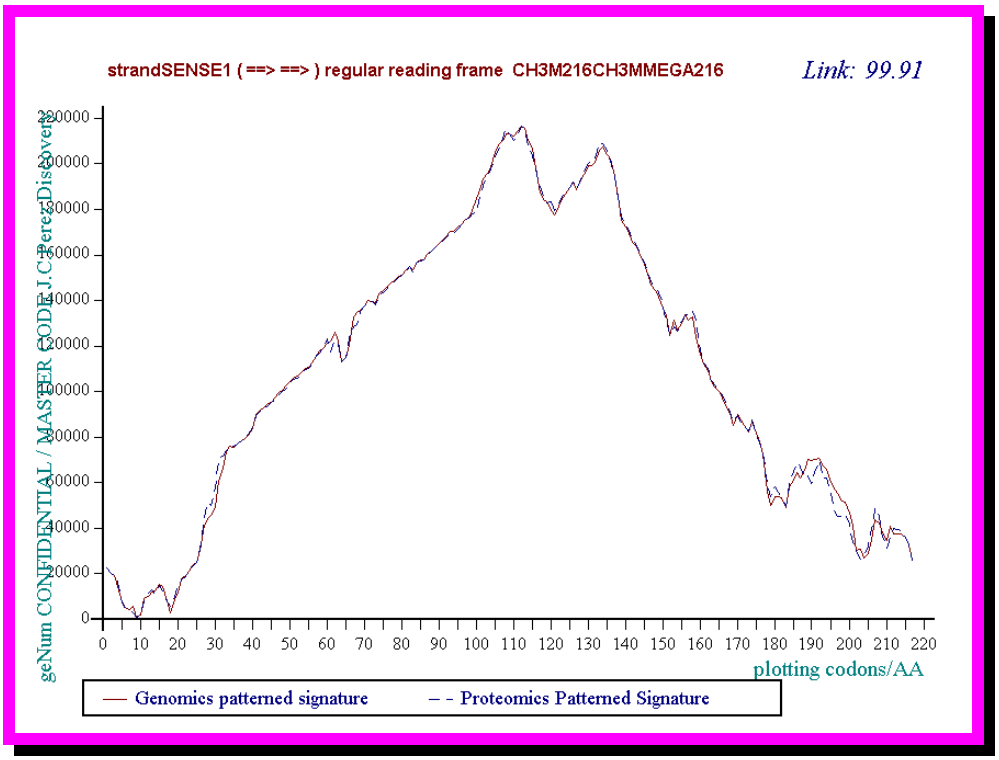
II-SUMMARY :

Mathematically, the "Universal Genomic Codes" discovery background is based alternatively on the three following mathematical worlds: Real Numbers Numerical World (ie. Atomic weights values), Integer Numbers World (ie. The Atomic Genetic Code described below), and a self-emerging LOGIC BINARY CODE 0/1, "Floor/Ceiling", "False/True"). The originality

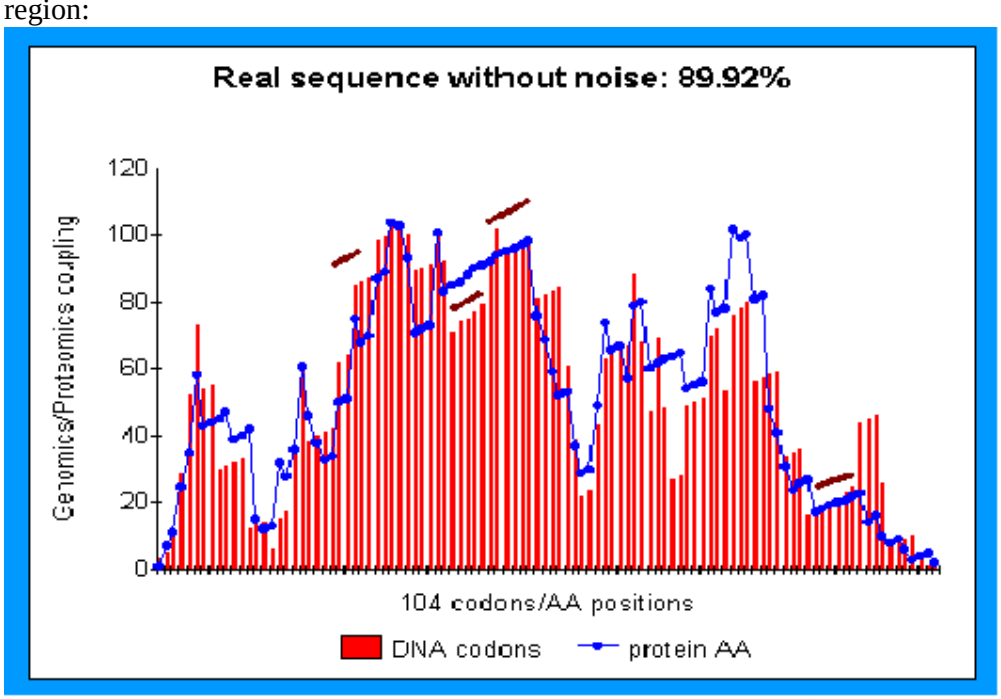
of this binary code it the following: that it is not an explicit and formal code but, on the contrary, a self-emerging code, this code is thus not an input but a consequence.
Globally, our discovery UNIFIES numerically the three GENETICS WORLDS: DNA, RNA and PROTEINS. This "great unification" unifies all genetic information from the six C O N H S P bio-atoms to the whole genome global level (ie U T C A G nucleotides, 20 amino acids, DNA or RNA strands, codons, anticodons, codons/anticodons couples, double-strand genomes etc...).
Particularly, we demonstrate 5 new unknown Biological Codes:
-the Atomic Genetic Code, where an INTEGER NUMBERS based common scale unifies DNA, RNA and proteins worlds starting from the atomic weights real numbers of the six C O N H S P bio-atoms. The SAME CODE emerges, simultaneously, by an analytical way, starting from the atomic composition of genetic compounds (Nucleotides, amino acids etc…).
-the Master Code, which proposes two highly correlated «numerical signatures » unifying any DNA sequence and its potential amino-acids translation. Curiously, the RNA numerical translation of the same sequence is always stabilized on the same "FIXED POINT", whatever the sequence, demonstrating the transitory fundamental nature of RNA (messenger information).
-the Binary Genomic Code (see below).
-the Undulatory Genomic Code, demonstrating the existence of Periodic Information Waves structuring all genomes.
-the Cytogenetic Banding Code, demonstrating, for the first way, that the well known experimental evidence of darker/lighter Bands characterizing Kariotypes is implicitely CODED in the TCAG DNA information. This code is a formal combination of both Binary and Undulatory Genomic Codes.

III-The BINARY LOGIC GENOMIC CODE Overview:

Mathematically, using NASH's vocabulary, the initial "GROUND LEVEL" background is provided by the exact atomic weights real numbers values of the six C O N H S P Bio-atoms. Then, we could compute atomic weights related to any DNA/RNA/protein compounds ie U T C A G nucleotides, the 20 amino acids, DNA or RNA codons etc... There are also REAL NUMBERS.
Then, using a simple common non linear projection formula (to be published), we obtain real numbers projections which aggregate them focusing around a scale of numerical "attractors" which are... INTEGER NUMBERS.
This translation provide now INTEGER NUMBERS based CODES which could be associated with any token as bio-atom, nucleotide, amino acid, DNA/RNA single-stranded or double-stranded codons or strains.
Then, coding the DNA sequence using this common law, we obtain integer numbers vectors associated with the DNA sequence simultaneously with the corresponding amino acids based translation (the unit elementary tokens are DNA codons and amino acids).
After a numerical integration-like discrete process, we obtain two curves or patterned signatures related to DNA and Proteomics worlds which are highly correlated (97% for the whole Human Genome), there is the "Master Code". Following an example of correlation between genomics image and proteomics image:



However, a fine analysis of the transitions from classifications between two consecutive neighbouring codons underlines a very curious phenomenon ***" in teeth of saw "*** which recall a bit FRACTALS-like patterns. For example, see the following zoom in a 312bps local region:



Then, analysing the texture (mathematical increase/decrease 1st degree derivates) of the Proteomics associated curve patterned signature, we could associate with each codon position "biospins" as following:

If the local codon position derivate is in increase state ➔ then Biospin=1,

If the local codon derivate derivate is in decrease state ➔ then Biospin=0.

Now, for any analysed sequence, we could compute the related balancing increase/decriease percentage related to the whole analysed sequence. These percentage values are real numbers in the range 0-100. Normally, the distribution of biospins percentages must be random, probably a Gauss-like distribution.
In fact, we obtain a very strange distribution as a "bath-tub"-like distribution: there appears, in ALL CASES, a binary distribution centered around two ATTRACTORS: one attractor, named "Floor-state attractor" is located around 29%.
The other second attractor, named "Ceiling-state attractor" is located about around 60%.
Please, see sample examples in both reports.
The following law is universal but restricted to GENOMIC DNA (and not EST transcripts, proteins etc...). ***We propose the following rule entitled "Genomic BINARY CODE law":***
For any sequence " seq " of genomic DNA, whatever its length, its position, and its nature, one can always associate, by applying the numerical algorithm described in (to be published),A BINARY CODE status, called « BioBit » such as:
BioBit (seq) = 0 = « Floor » state = « FALSE » if %(seq) neighbouring attractor 29%.
BioBit (seq) = 1 = « Ceiling » state = « TRUE » if %(seq) neighbouring attractor 60%.
We validated and checked this universal law on the totality of the genomes known to date, and, more particularly, on the whole human genome which we studied independantly on three embedded scales: contiguous segments of 10000bases, 100000bases and 1million of bases.

IV-RESULTS :

We demonstrate now this "Natural Hierarchical Introspective Logics" on a randomly selected region within the Draft Human Genome sequence. This genomic studied region is located between 130000000 and 131024000 positions within the human chromosome5.
Some regions are undefined ("N" undefined bases or "GAPS").

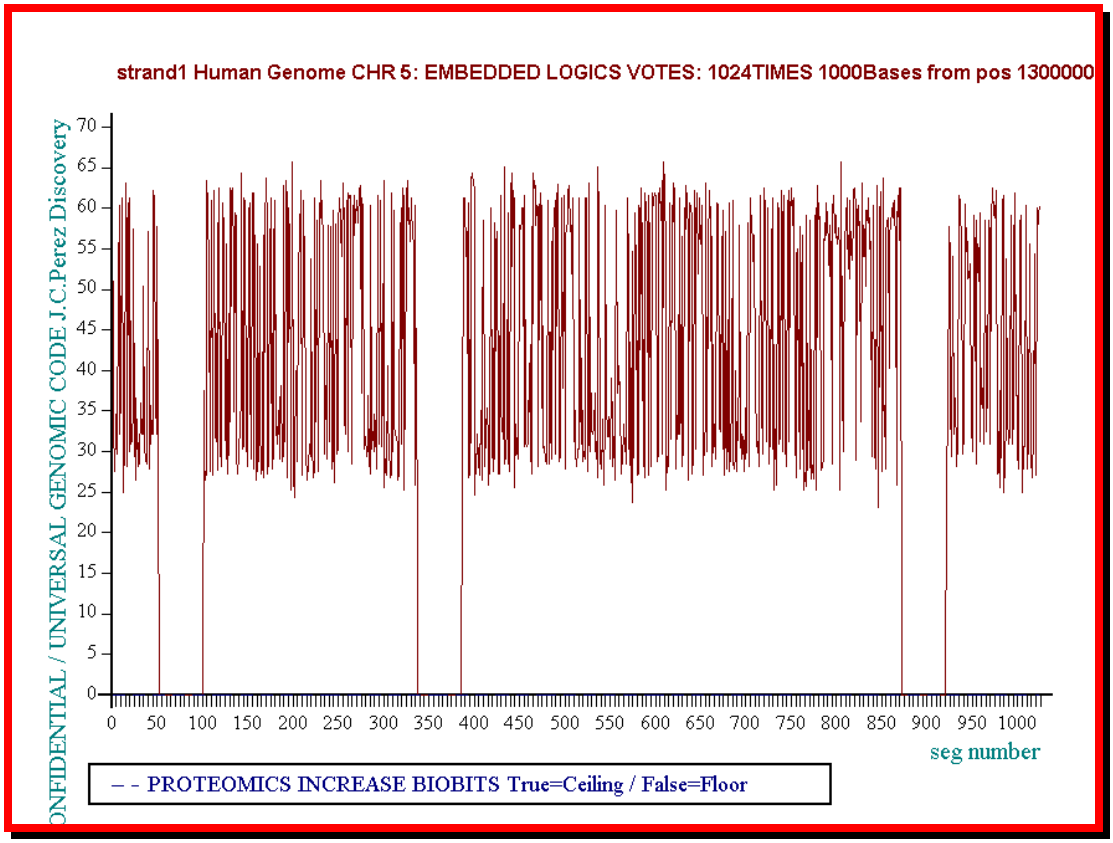
We run 11 independant embedded analyses:
-1024 contiguous DNA segments of 1000bases.
- 512 contiguous DNA segments of 2000bases.
- 256 contiguous DNA segments of 4000bases.
- 128 contiguous DNA segments of 8000bases.
- 64 contiguous DNA segments of 16000bases.
- 32 contiguous DNA segments of 32000bases.
- 16 contiguous DNA segments of 64000bases.
- 8 contiguous DNA segments of 128000bases.
- 4 contiguous DNA segments of 256000bases.
- 2 contiguous DNA segments of 512000bases.
- 1 unique DNA segment of 1024000bases.

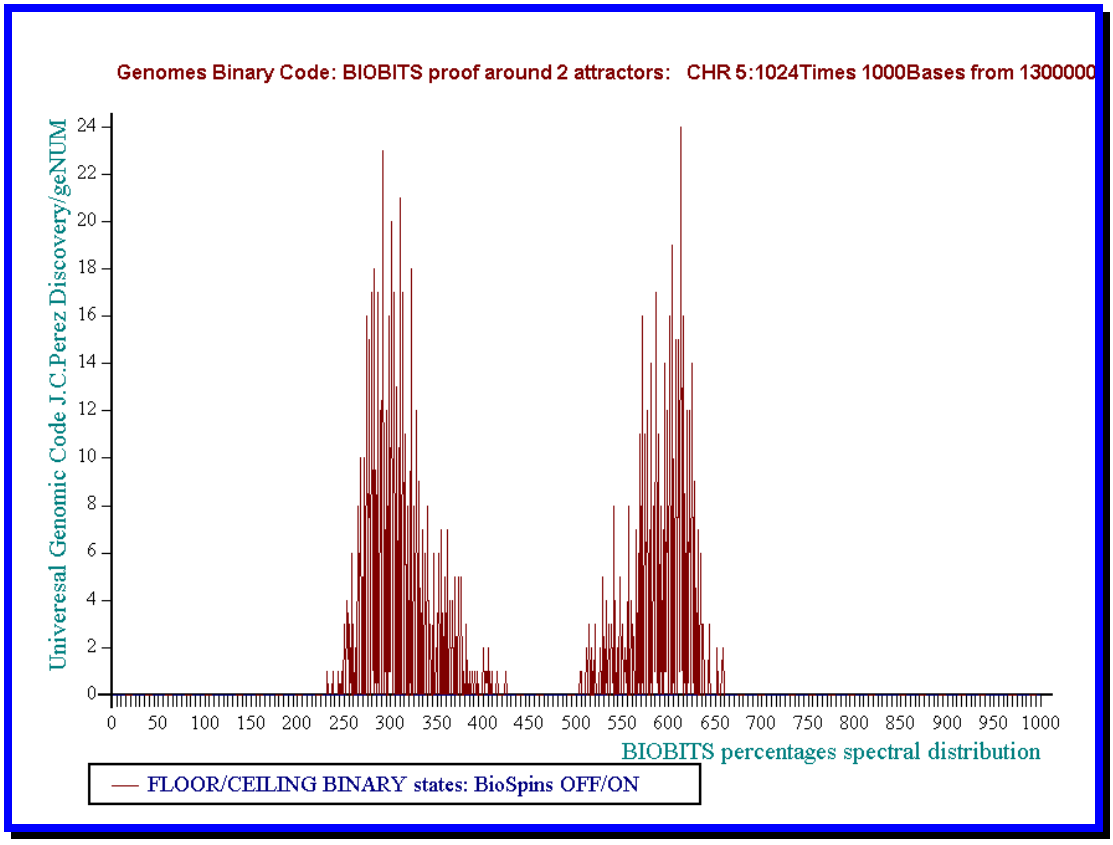
In the following table, we resume, for the 11 independant analyses:
-the numbers of elementary BioBits decisions: exp in line 1: 468 "Floor states" and 409 "Ceiling states", the total correspond to 877 segments, the remaining are GAP segments.
-the average values of Floor and Ceiling percentages: exp in line 1: 31% for "Floor attractor" and 59% for "Ceiling attractor".
-the LOCAL LEVEL VOTE DECISION: ***exp in line1, the Floor state (468) is majority*** then ***the local level decision is "Floor"=FALSE***.

11 Embedded levels	Floor/FALSE/0 states: Average Floor %	Ceiling/TRUE/1 states: Average Ceiling %	Level Vote Decision
First Level 1024 times 1000bases	468 floor level 31%	409 ceiling level 59%	Floor=FALSE
Second Level 512 times 2000bases	239 floor level 30%	201 ceiling level 60%	Floor=FALSE
Third Level 256 times 4000bases	123 floor level 29%	98 ceiling level 60%	Floor=FALSE
Fourth Level 128 times 8000bases	65 floor level 29%	47 ceiling level 61%	Floor=FALSE
Fifth Level 64 times 16000bases	34 floor level 29%	24 ceiling level 61%	Floor=FALSE
Sixth Level 32 times 32000bases	19 floor level 28%	11 ceiling level 61%	Floor=FALSE
Seventh Level 16 times 64000bases	8 (*) floor level 28%	8 (*) ceiling level 61%	UNDEFINED (*)
Eighth Level 8 times 128000bases	7 floor level 27%	1 ceiling level 60%	Floor=FALSE
Ninth Level 4 times 256000bases	3 floor level 27%	1 ceiling level 61%	Floor=FALSE
Tenth Level 2 times 512000bases	2 floor level 27%	0 ceiling level 0%	Floor=FALSE
Final Eleven Level 1 time 1024000bases	1 floor level 27%	0 ceiling level 0%	Floor=FALSE

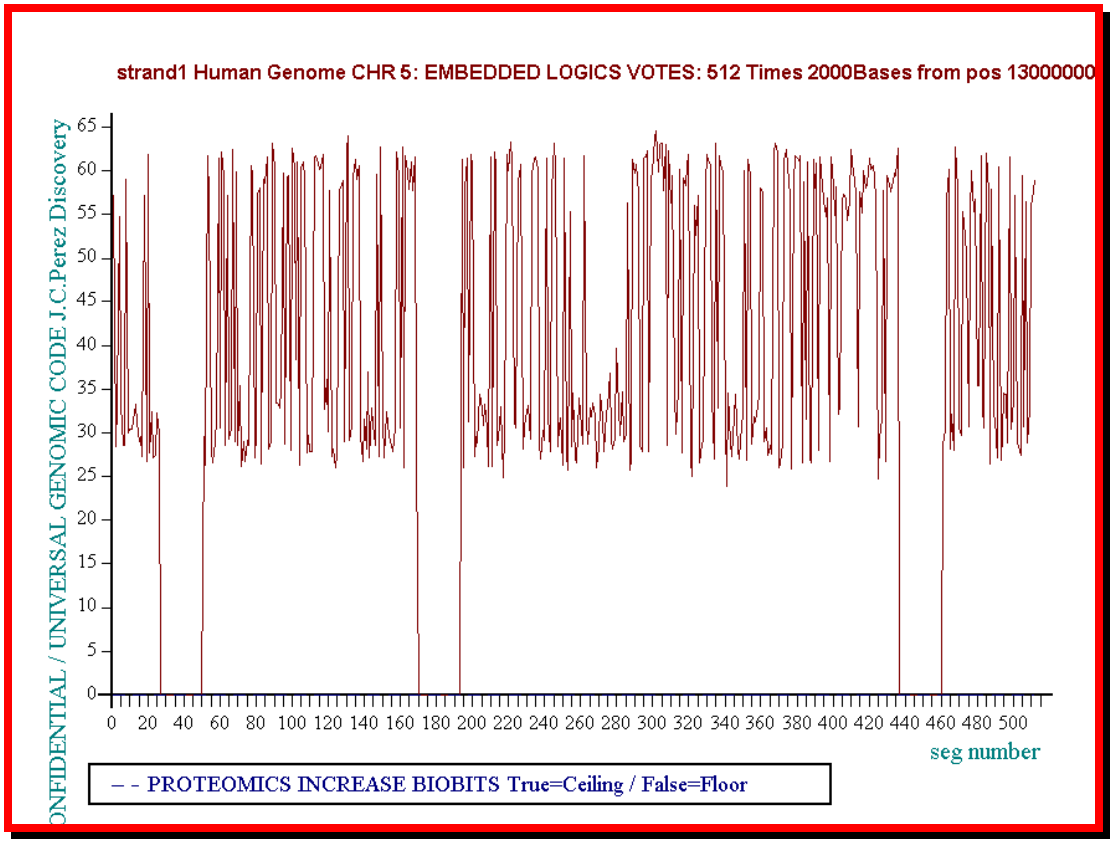
(*) Nota: 4 Ceiling states are not significant (they include large gaps undefined DNA regions). In the following couples of graphics we demonstrate the SCALE INVARIANCE and the EMBEDDED LOGICAL VOTE process.

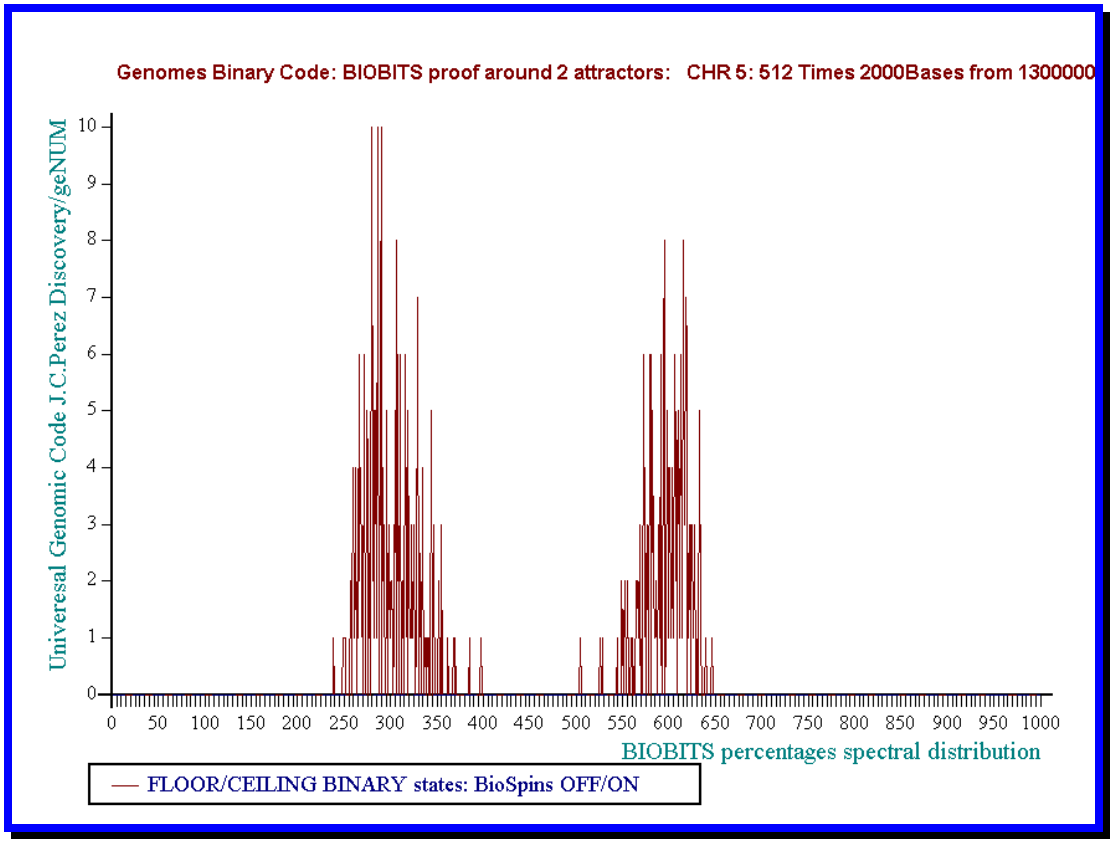
➔ Level 1: ***1024 times 1000*** bases... **Concensus Decision "VOTE" = Floor = "False"**



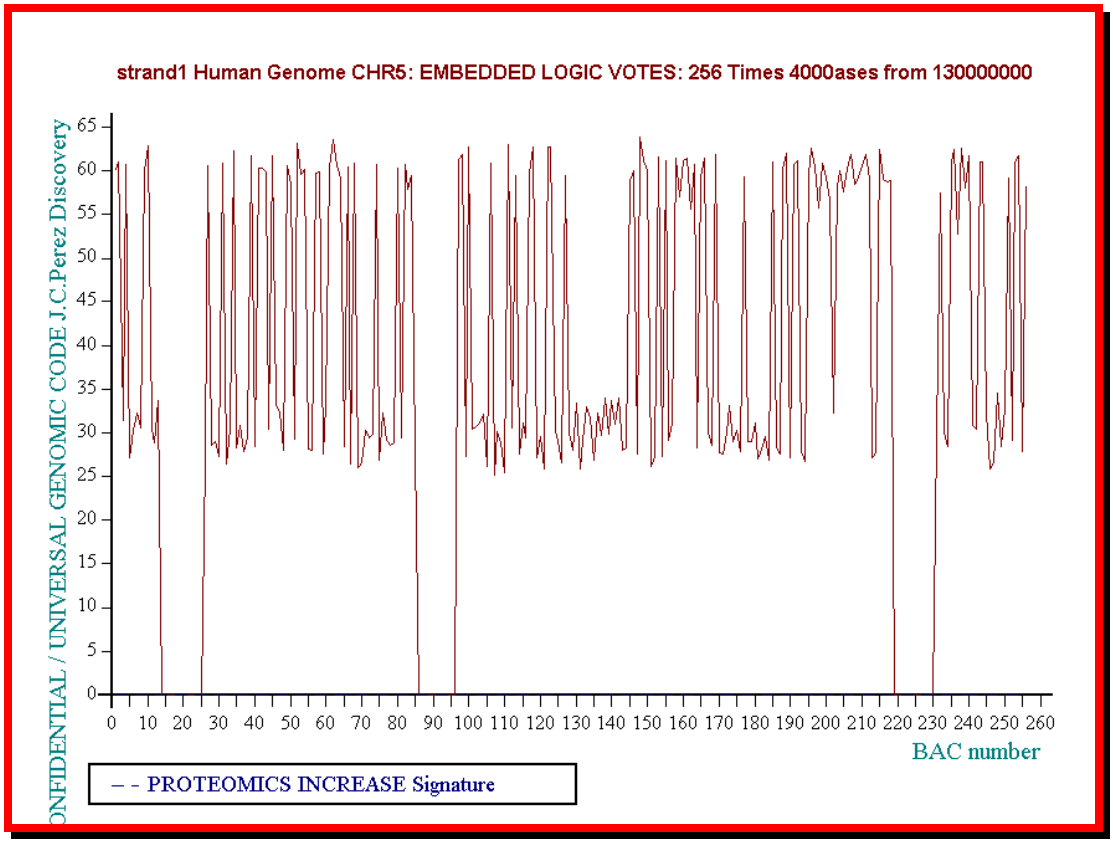


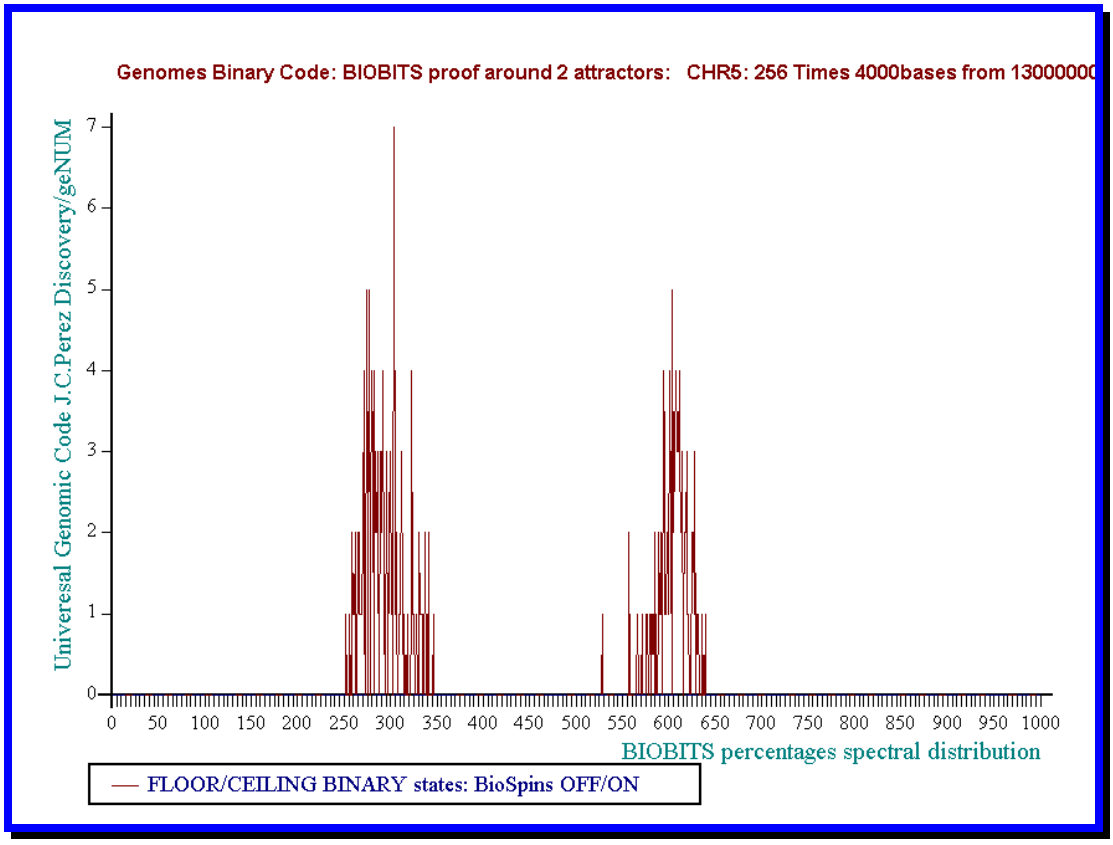
➔ Level 2: ***512 times 2000*** bases... **Concensus Decision "VOTE" = Floor = "False"**



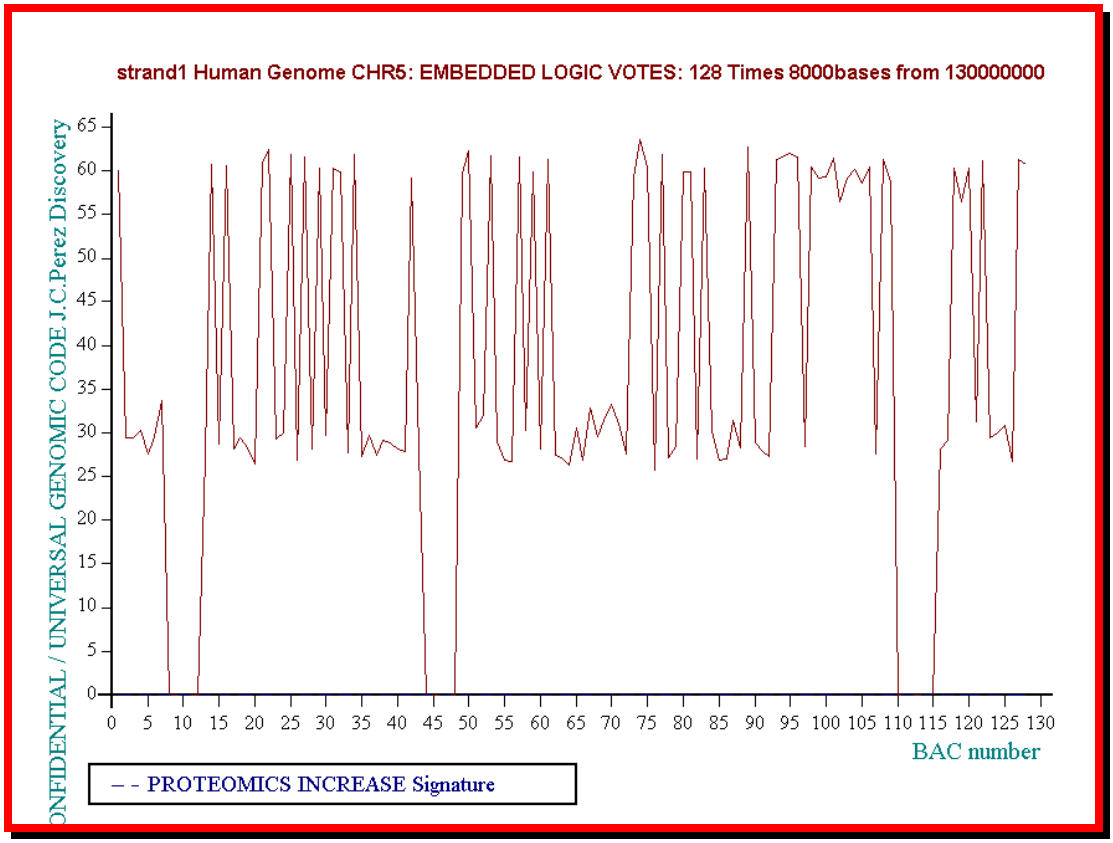


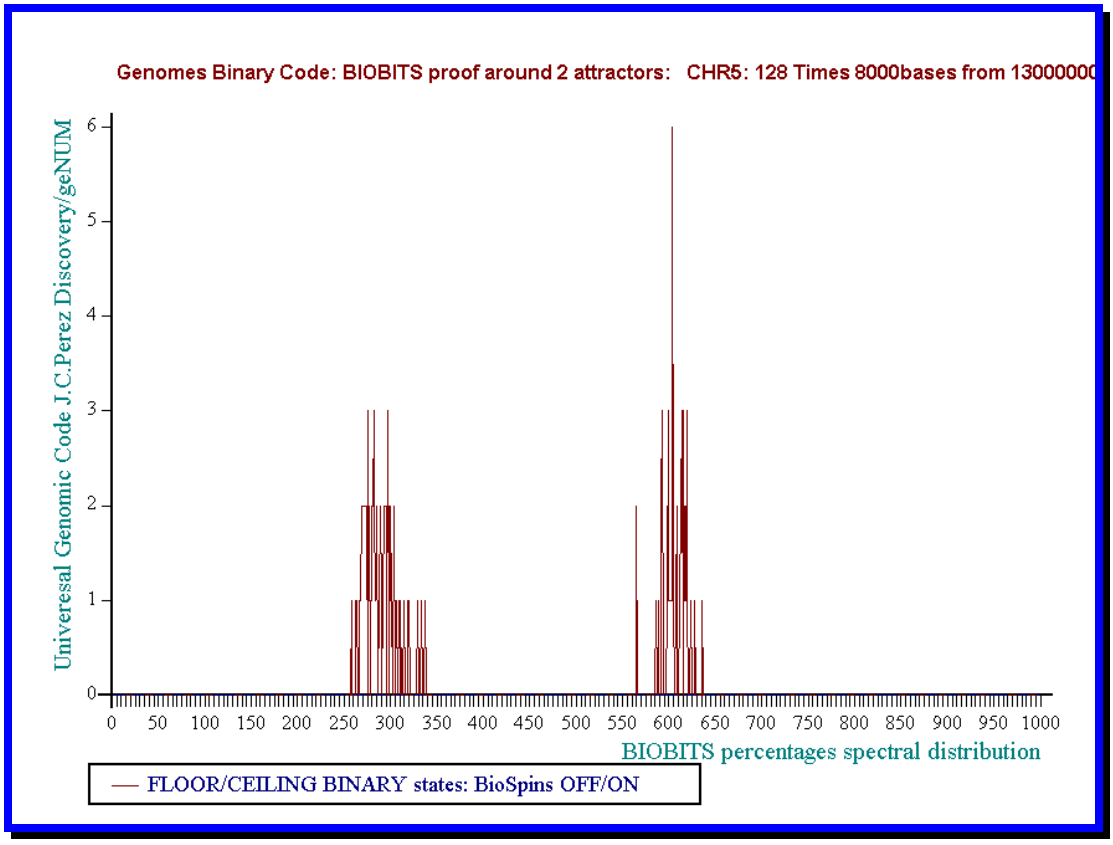
➔ Level 3: ***256 times 4000*** bases... **Concensus Decision "VOTE" = Floor = "False"**



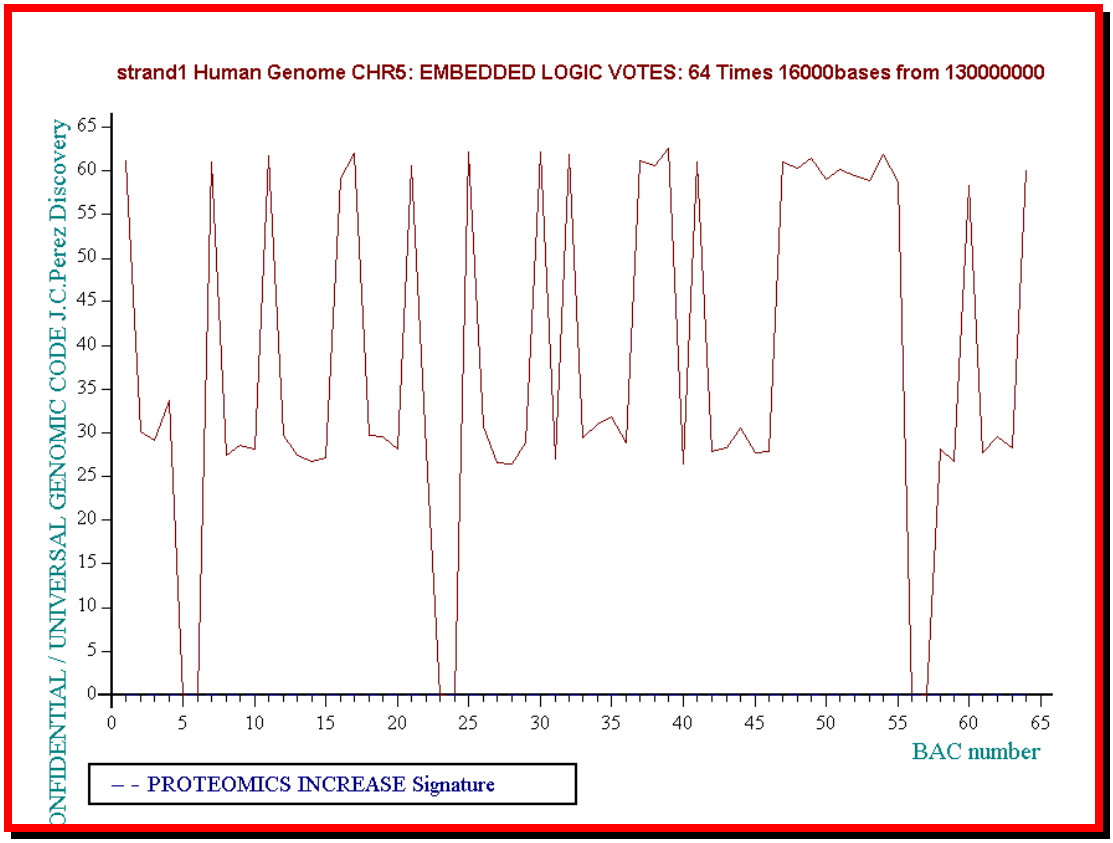


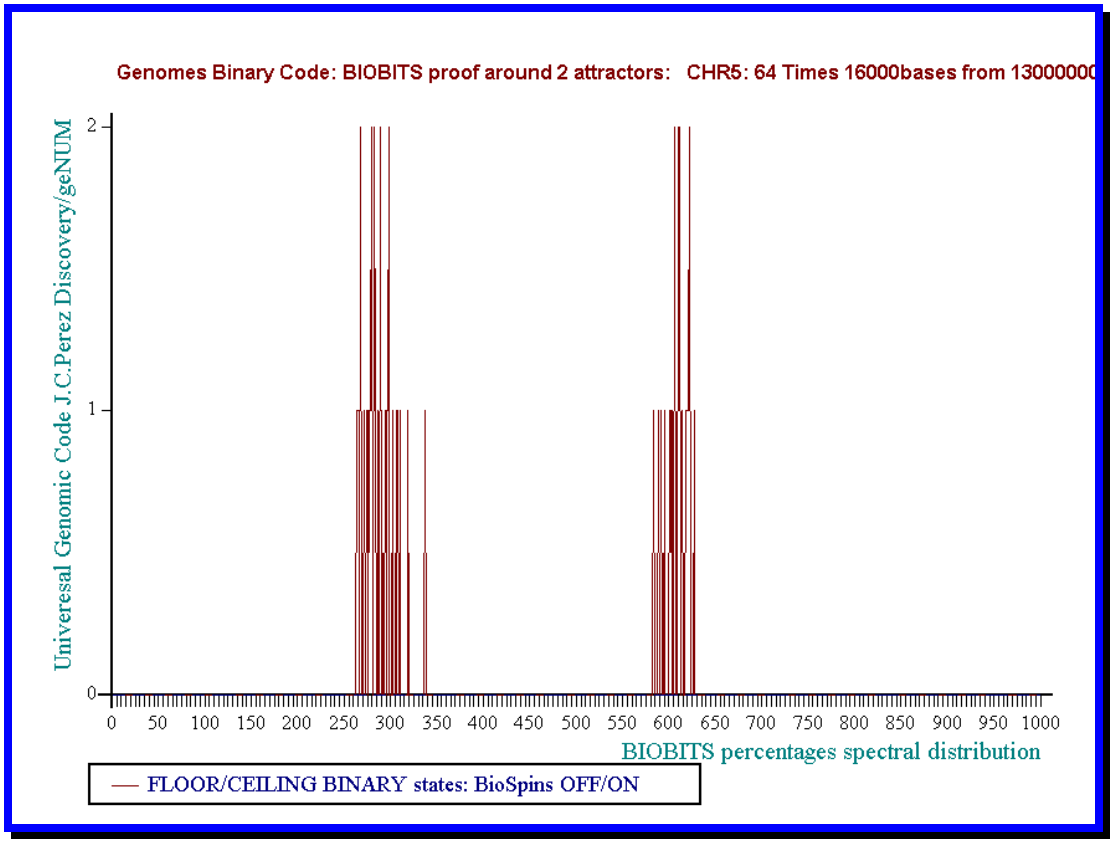
➔ Level 4: ***128 times 8000*** bases... **Concensus Decision "VOTE" = Floor = "False"**



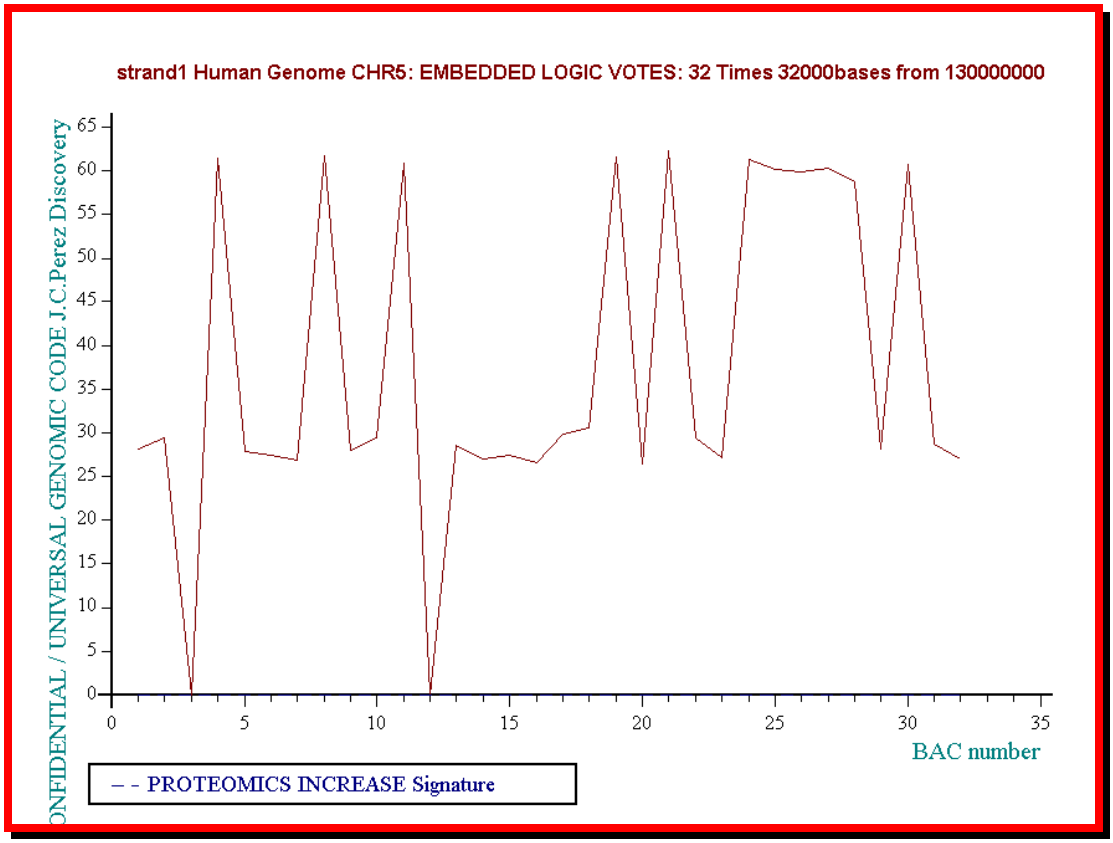


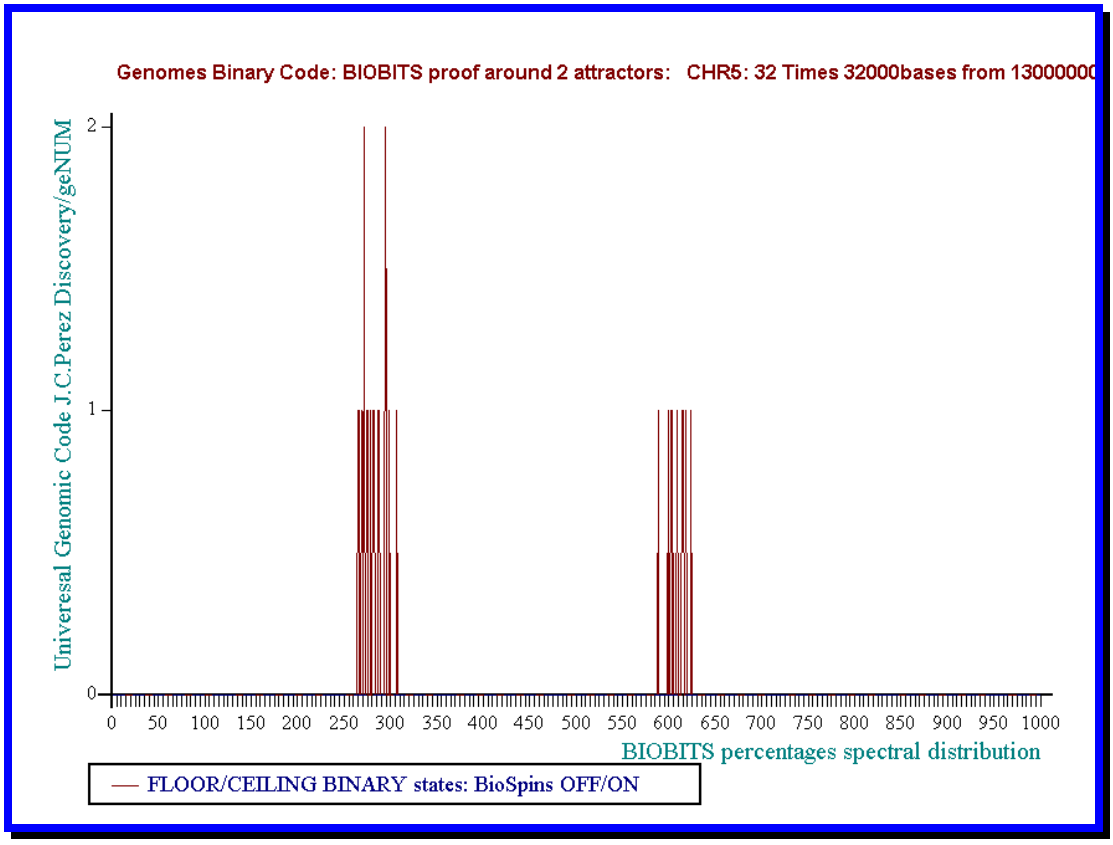
➔ Level 5: ***64 times 16000*** bases... **Concensus Decision "VOTE" = Floor = "False"**



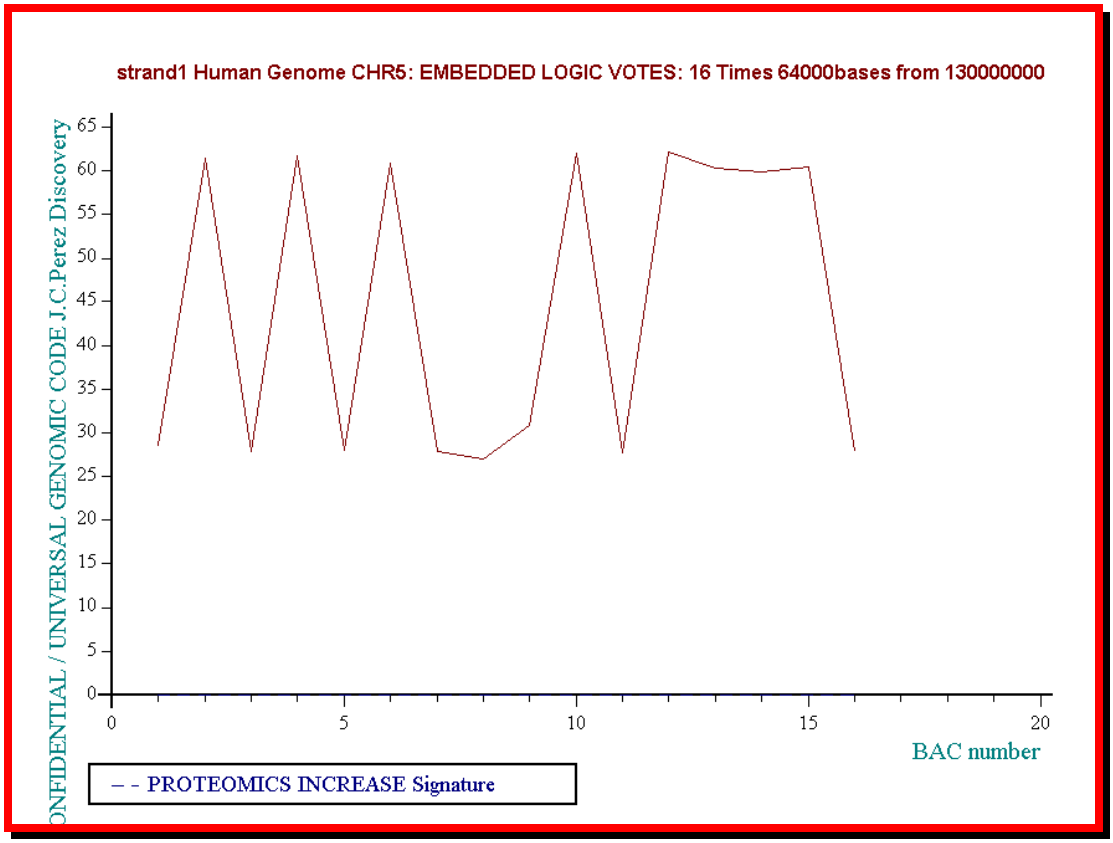


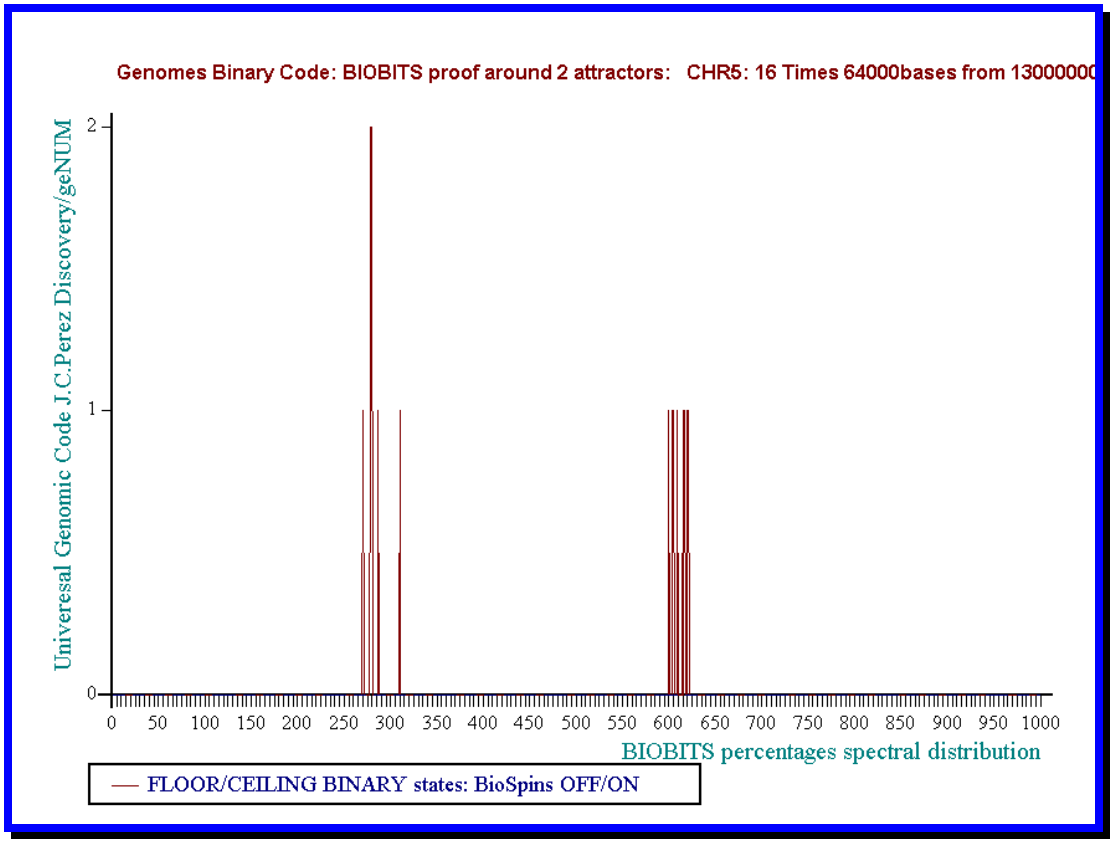
➔ Level 6: ***32 times 32000*** bases... **Concensus Decision "VOTE" = Floor = "False"**



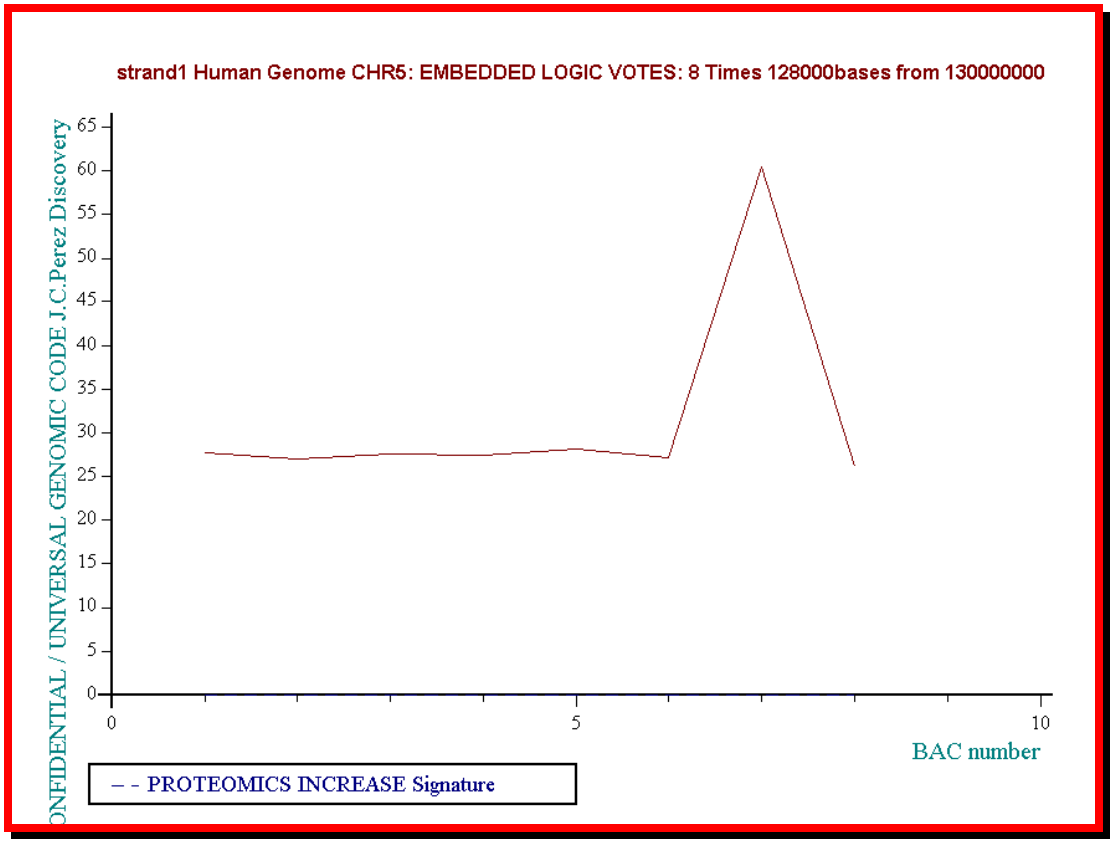


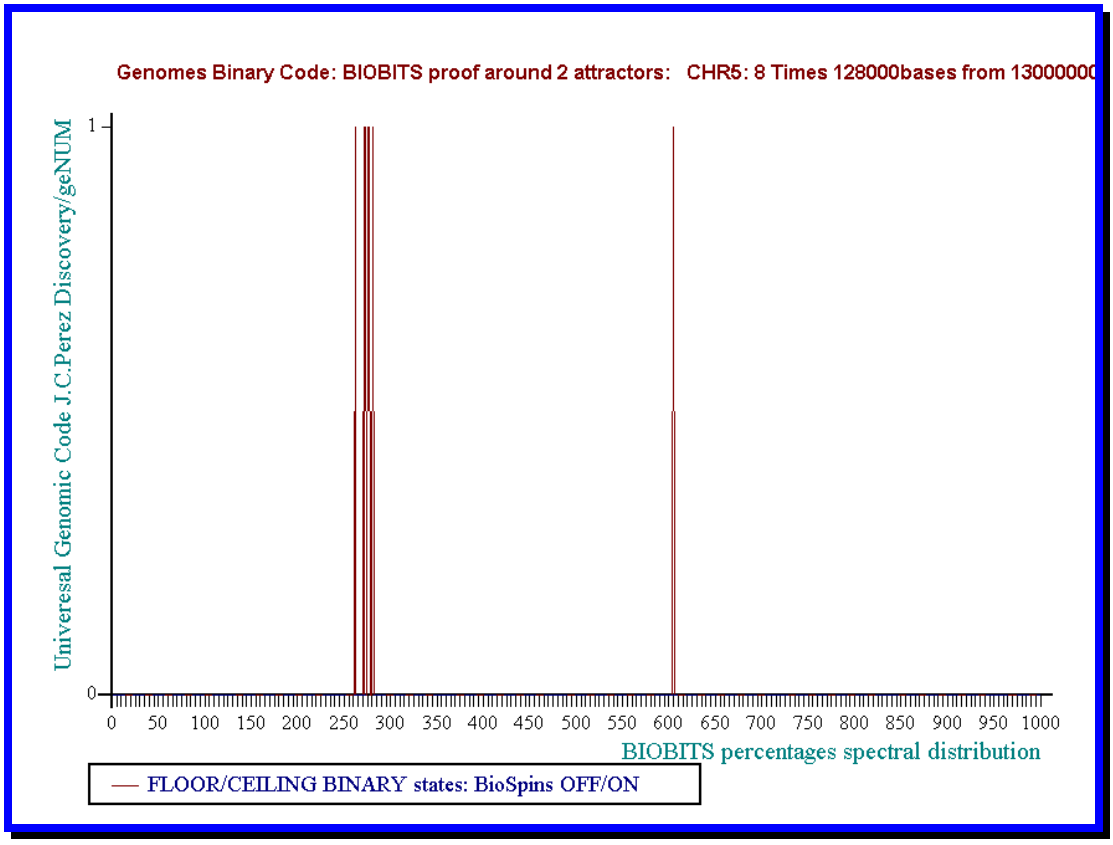
➔ Level 7: ***16 times 64000*** bases... **Concensus Decision "VOTE" = "Undefined"**



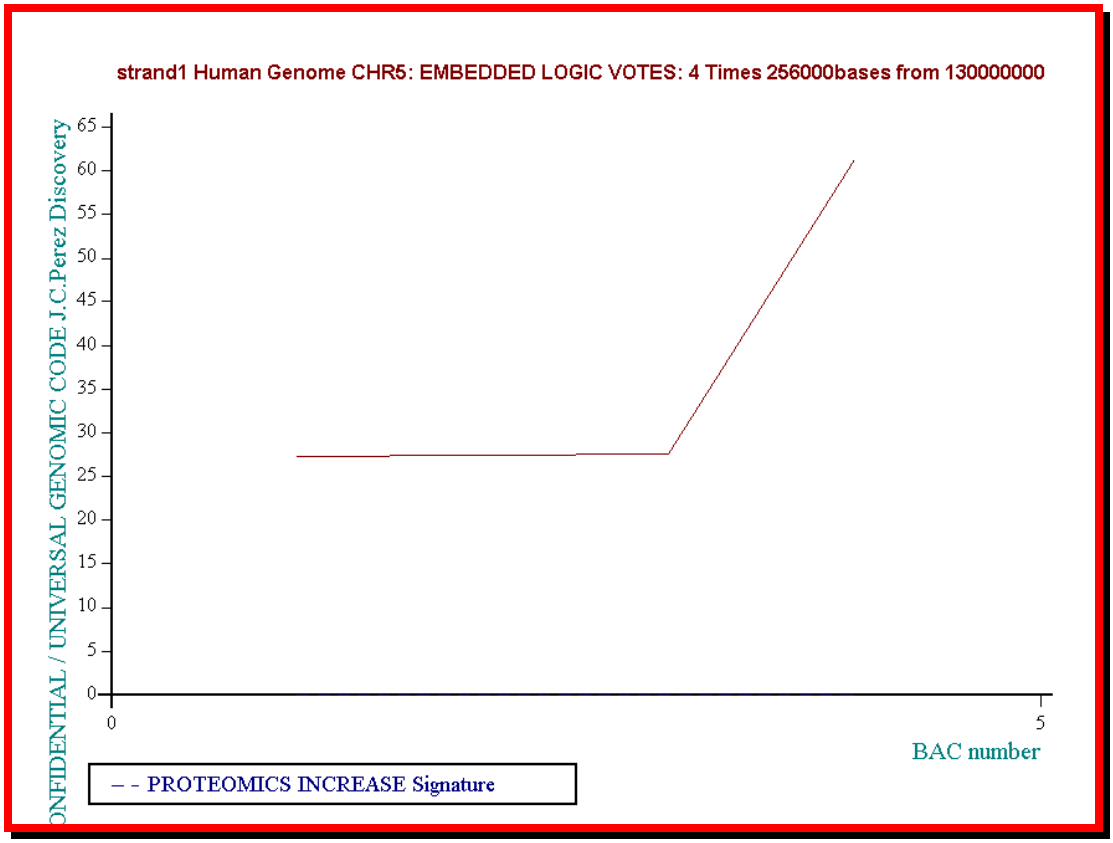


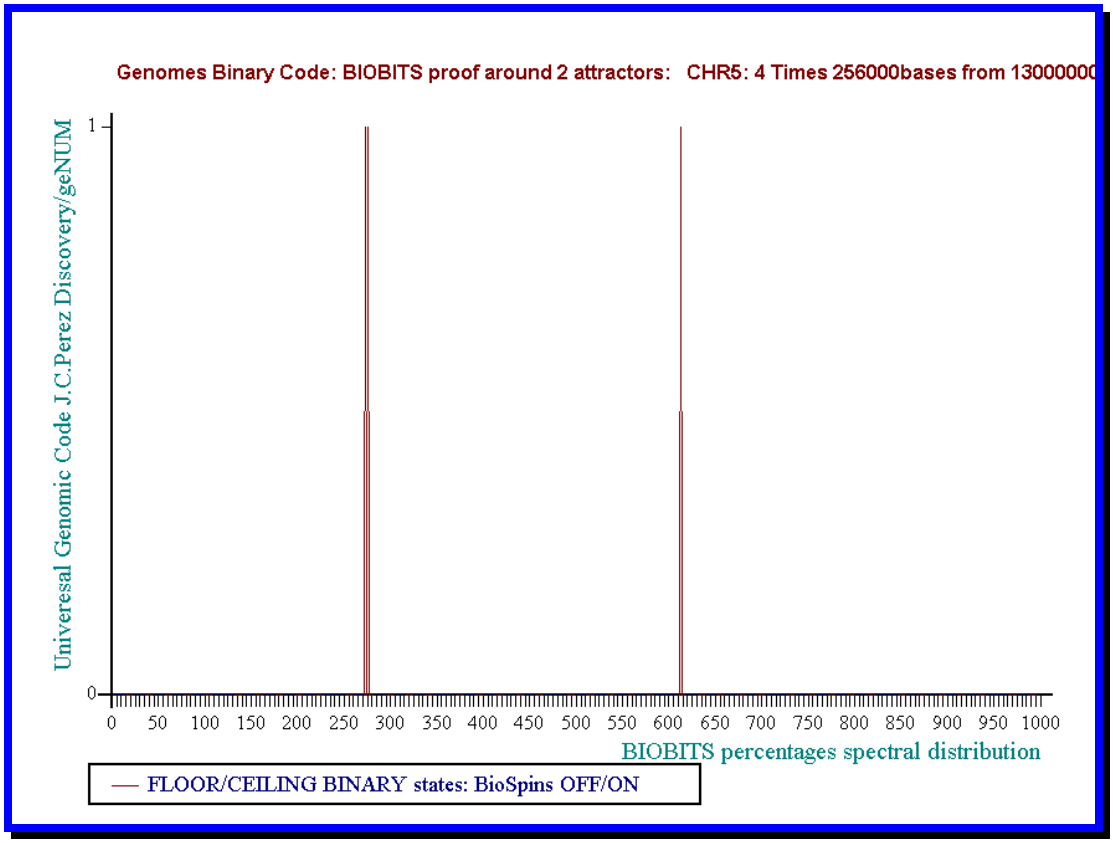
➔ Level 8: ***8 times 128000*** bases... **Concensus Decision "VOTE" = Floor = "False"**



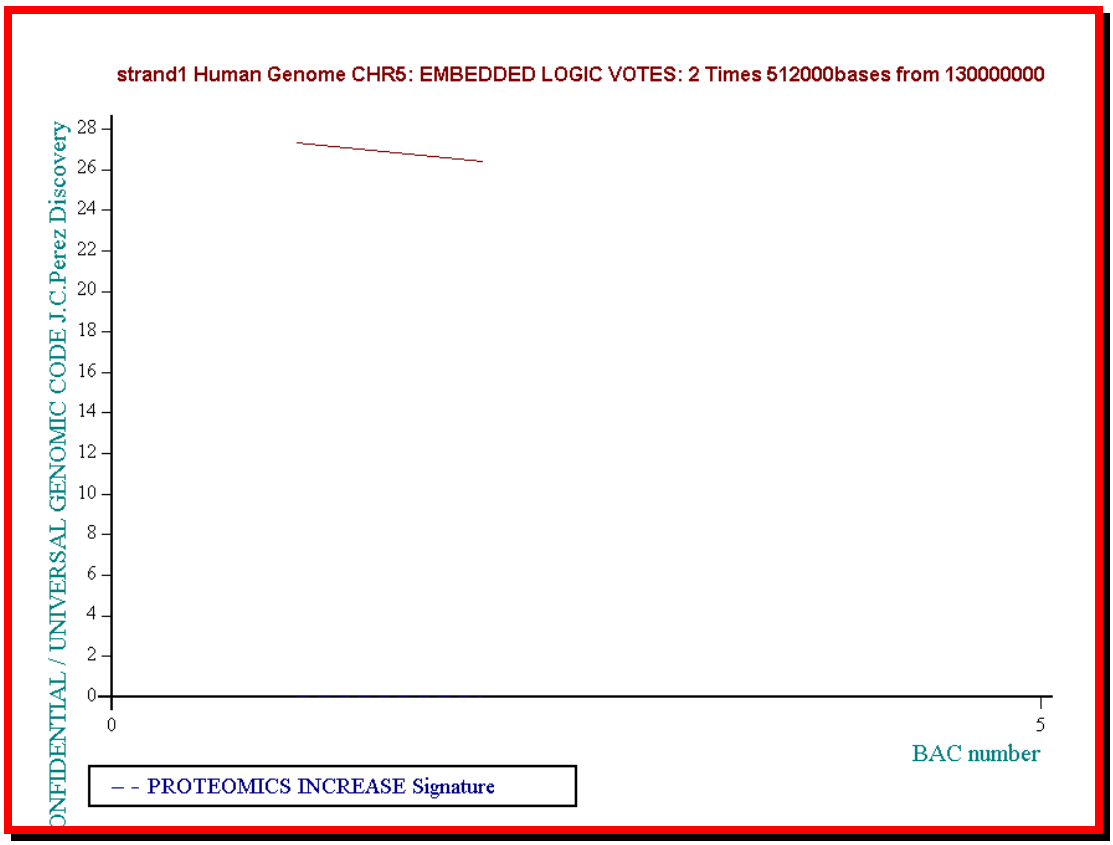


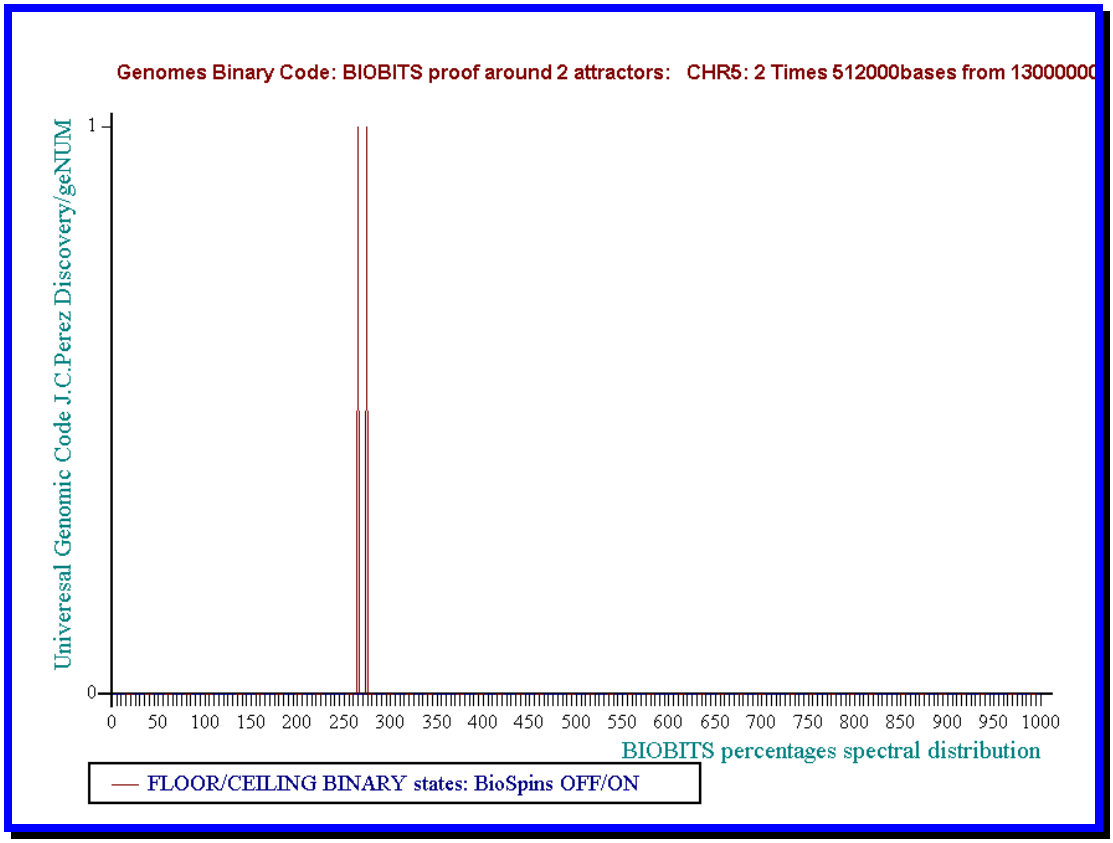
➔ Level 9: ***4 times 256000*** bases... **Concensus Decision "VOTE" = Floor = "False"**



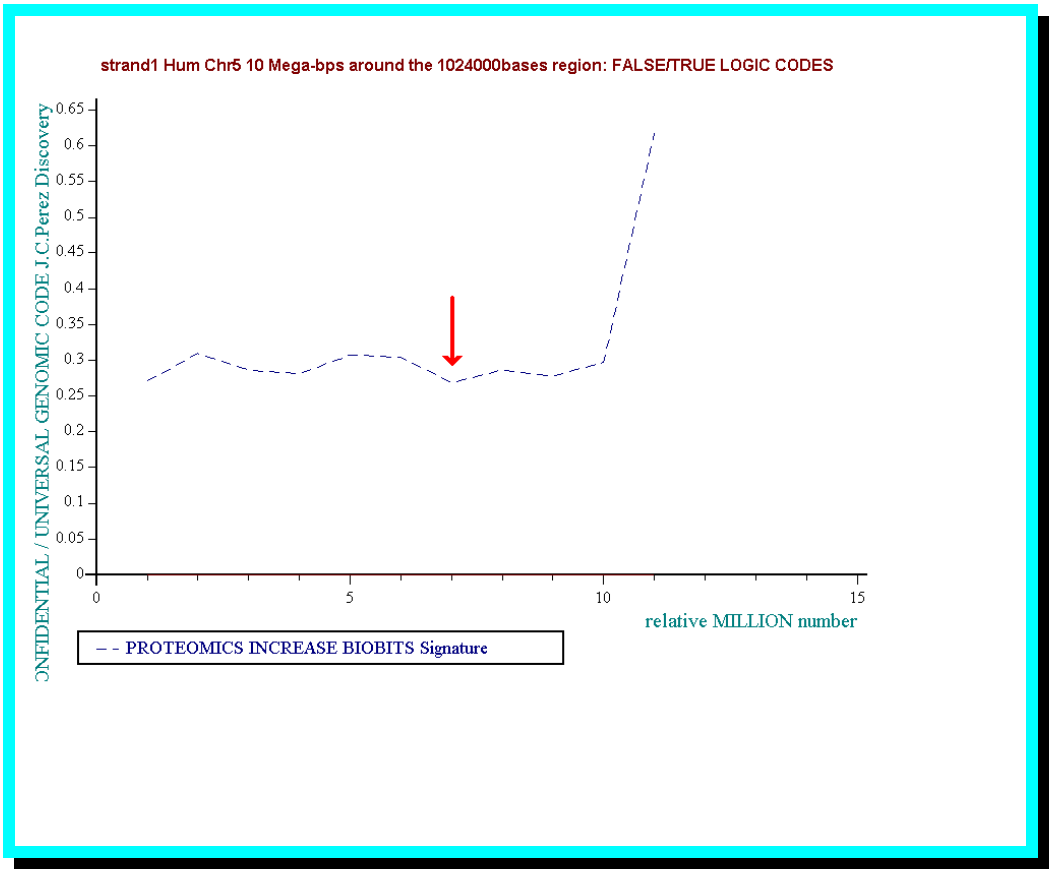


➔ Level 10: ***2 times 512000*** bases... **Concensus Decision "VOTE" = Floor = "False"**


strand1 Human Genome CHR5: EMBEDDED LOGIC VOTES: 2 Times 512000bases from 130000000
ONFIDENTIAL / UNIVERSAL GENOMIC CODE J.C.Perez Discovery
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
0
5
BAC number
– - PROTEOMICS INCREASE Signature


Genomes Binary Code: BIOBITS proof around 2 attractors: CHR5: 2 Times 512000bases from 13000000
Univeresal Genomic Code J.C.Perez Discovery/geNUM
1
0
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1000
BIOBITS percentages spectral distribution
— FLOOR/CEILING BINARY states: BioSpins OFF/ON

In the following graphic, we summarize the genomic area around the 1024000bases DNA sequence analysed. This 10 millions lenght sequence was analysed spliting it in 10 “ONE MILLION regions”. The red arrow localizes the studied 1024000 studied region. Then, the area is globally at “Floor” state, with a “Ceiling” state transition at the end (see on the right).



V-DISCUSSION :

As Professor NASH in " Hierarchical Introspective Logics “proposes it:
On the one hand, the concept of incompleteness will be able to evolve with the levels of human knowledge. In addition, the discovery of new natural laws will be able to make evolve the approach of this problem “.../... But the history of human progress in science and mathematics reveals that observation of the phenomena of Nature has always played a large rôle.../...” (by J. F. NASH in the above paper).
In other hand, the concept of LOGIC suggested here is radically new because it acts of an *self-emerging logic*, output of a self-organized multi-levels embedded process of which the roots (" ground level ") are at the basic level of the “average atomic masses of the 6 DNA CONHSP bio-atoms”, therefore in a world of real numbers.
Lastly, the concept of hierarchy is not discrete but completely continue in an infinity of embedded levels. Thus, the choice of 11 levels in the example suggested is arbitrary, one could also have chosen thousands of embedded levels... We show thus that the human genome (as all the other genomes) is the source of an omnipresent logical binary language which appears to be invariant on all the scales. This code is not explicit and formal but self-emergent as the " output " of a complex genetic system. This code is embedded in a self-referred infinity of VOTE-like level. Perhaps it could give new ways and tracks to understand the “Natural Hierarchica Introspective Logics” decision making process.

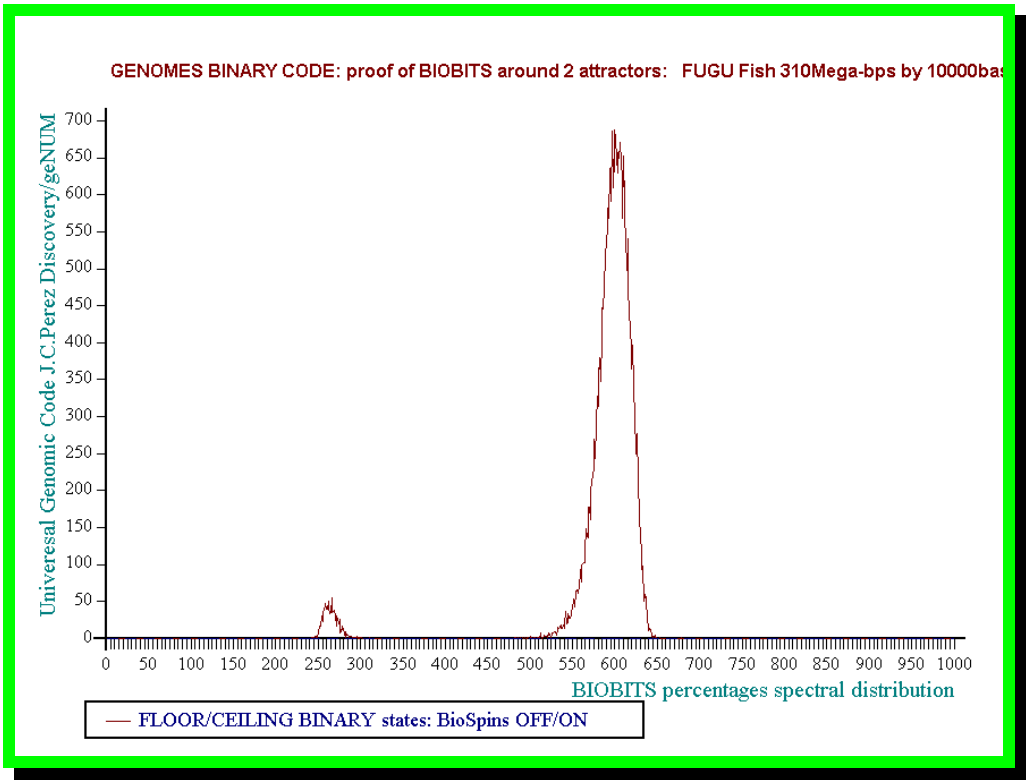
VI-GENERALIZATION to the WHOLE HUMAN GENOME:

Now, we show that the discuted self-organized binary code could be generalized to ALL genomes and particularly to the whole Human Genome...
In this section, we must talk a bit about the ***"Golden Ratio":***
Rare papers provide trace of golden ratio within DNA... Are really DNA, genes and whole genomes controled by this strange universal constant? In 1991 the author find presence of Fibonacci numbers then of golden ratio between related proportions of TCAG nuceotides within the DNA of genes[1] then confirmed by a book 1997[2]. In 2007, in the bulletin of Mathematical Biology, Yamagishi and Shimabukuro established an interesting connection between nucleotide frequencies in human single-stranded DNA and the famous Fibonacci's numbers [3].

VARIOUS LONG GENOMES:

The discussed discovery was improved on lots of genomes: bacterias, archaea, eucharyotes etc... In all casess, with nuance like here (FUGU fish) the evidence of the 2 attractors self-organization remains.

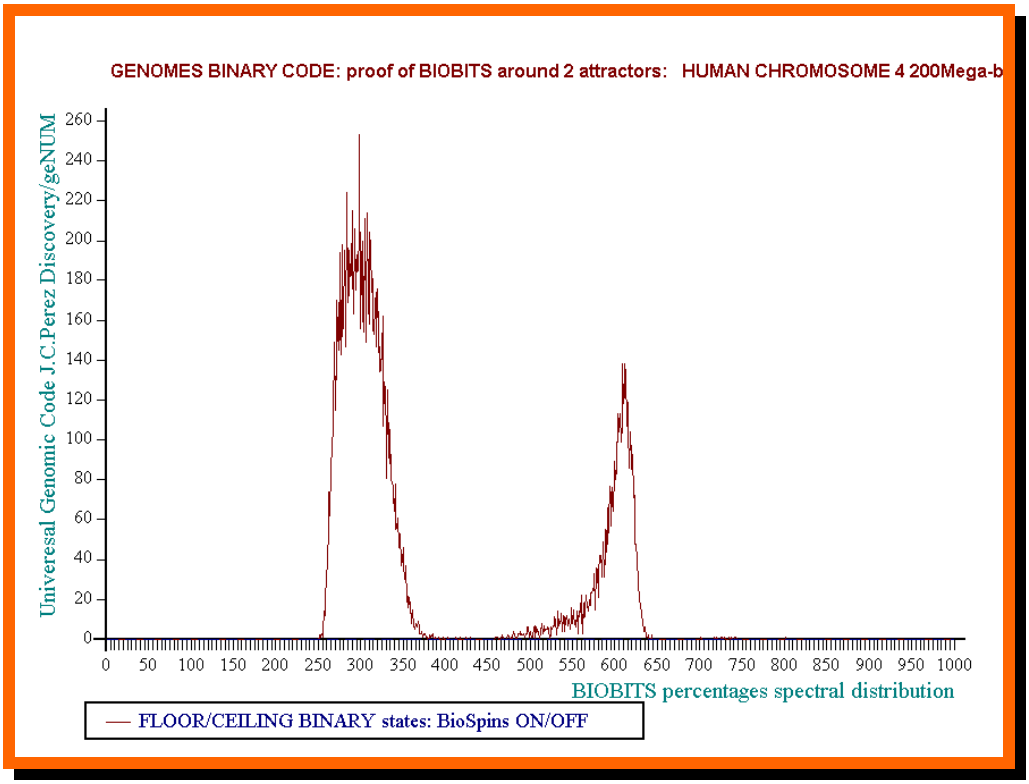
➔ SPECTRUM Binary Code FUGU Fish genome (310Megas by 10000bps):



[1] J.C. Perez - "Chaos DNA and Neuro-computers : a golden link / The hidden language of genes, global language and ordre in the human genome", in Speculations in Science and Technology, vol 14 number 4 1991, ISSN 0155-7785.
[2] Perez, Jean-claude (1997). *L'ADN décrypté*. Embourg (Belgium): Marco Pietteur. ISBN 2-87211-017-8.
[3] Yamagishi M. E. B. and Shimabukuro A. I. (2007) Nucleotide Frequencies in Human Genome and Fibonacci Numbers. Bulletin of Mathematical Biology

THE WHOLE HUMAN GENOME:

➔ SPECTRUM Binary Code of HUMAN chr4 (200Megas by 10000bps):



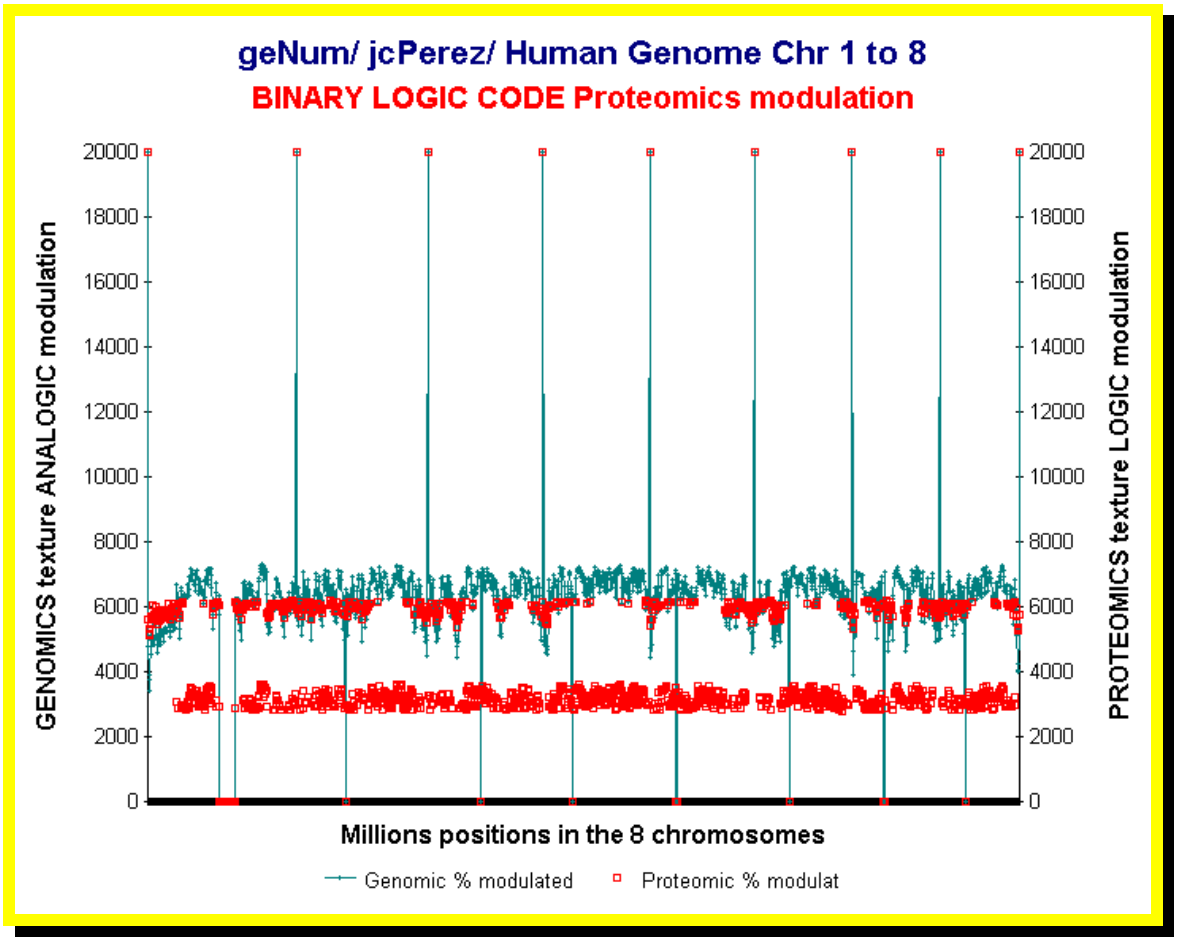
« BITS » overlapping the complete 3billions bases-pairs Human genome: The BINARY 0/1 CODE.

Notes ralated to both next figures :

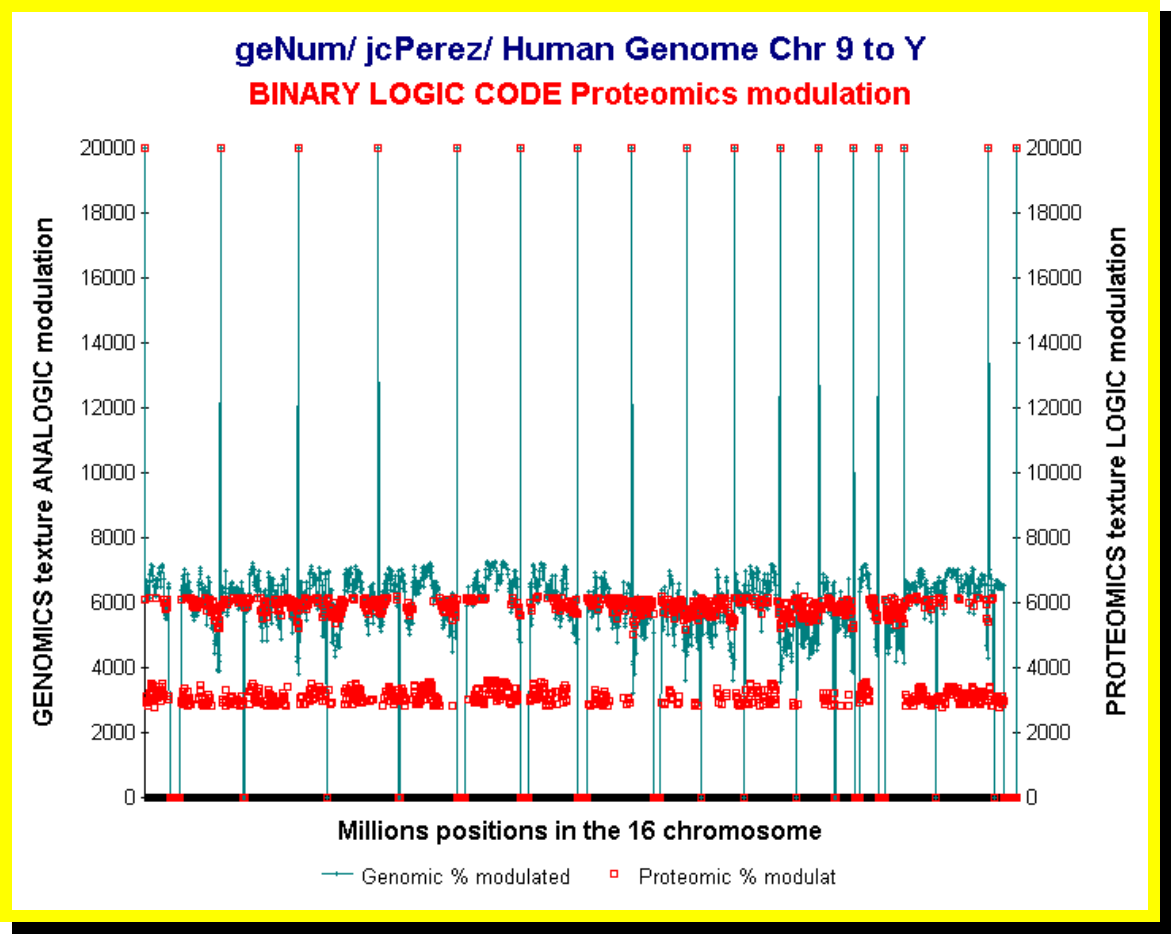
- in each graph the unit element in X-coordinates is the million bases pair (on the whole: 3266 units representing 3.266 billion bases. Among them, 3075 million relates to significant areas, the 191 remaining million relate to areas GAP (unspecified “N” bases), in particular the centromères of the chromosomes. These last areas are located by the small red points on the x-axis.
- the green vertical bars delimit various chromosomes frontiers and centromeres.
- the two represented indicators come from the measurement of “textures-like”, on the one hand on curves GENOMICS, on the other hand on curves PROTEOMICS relating to each million bases analyzed separately.
- Although ALL these couples of curves GENOMICS and PROTEOMICS of the MASTER CODE are very strongly correllated and quasi-superposable (96.63% on average on the whole human genome) and independently of the 3 codons reading frames and the 4 logical combinations of the possible directions of readings of the 2 DNA strands, it is observed that:
- if these couples of patterned signatures of the MASTER CODES UNIFY by their FORMS, on the contrary, and it is an astonishing fact there, they are DIFFERENCIED by their TEXTURES: the texture of curves GENOMICS "is modulated in an ANALOGICAL way" around an average value close to 60% (graduation 6000).

- On the contrary, the texture of curves PROTEOMICS (although calculated in an identical way, according to same rules', and on an quasi-identical curve), it, " is modulated in a BINARY LOGICAL way " oscillating between two attractors whose average values are respectively of: FLOOR=30% on average, then CEILING=60% on average. The average ratio between these two binary states is very close to TWO (1.97). The cloud of the red points illustrates this binary 0/1 variable "FLOOR/CEILING ".We showed that the sequence of these transitions from states 0/1 is in direct connection with famous CYTOGENETICS BANDS of the CHROMOSOMES: the KARIOTYPE of the human genome…
- **SUMMARIZING and as a synthesis: A mathematical reality, having taken its roots on the level of the exact atomic weights of the 6 bio-atoms C (Carbon), O (Oxygen), N (nitrogenizes), H (Hydrogen), S (Sulphur) and P (Phosphorus) will have thus led us until the discovery of a global structure to the level of the whole genomes. This structure, the " MASTER CODES " underlines not less than three levels of languages of which the emergence of WAVES and a BINARY LOGIC. We show finally that these Languages are completely coupled with the EXPERIMENTAL OBSERVATION of the dark and clear Bands of the KARYOTYPES which Master Codes allows predictions…**

The 8 Chromosomes 1 to 8 of the whole draft HUMAN GENOME…



The 16 remaining Chromosomes 9 to Y of the draft HUMAN GENOME…



ABOUT VALUES OF THE 2 ATTRACTORS...

The Whole Human Genome Binary Code:

The whole Human Genome is controled by two BINARY CODES ATTRACTORS which provide a kind of self-organized bistable binary code ... like in computers! With the central following difference:
-the binary code within computers was invented artificially by humans...
-the binary code of Life has “emerged” spontaneously ...
perhaps by self-organization... or... perhaps by “another unknown” action?
MEANWHILE, there are facts:
-The ratio between both bistable states is exactly equal to “2” (the space between two consecutives octaves in Music...)
-The Top state is exactly matching with a GOLDEN RATIO...
-The Bottom state is also exactly related to Golden Ratio...
“Top” level = $\varphi = 1 / \Phi$
“Bottom” level = $\varphi / 2 = 1 / 2\Phi$
Top / Bottom = 2 ... Where Φ is the « Golden Ratio”...