[Frontiers in Bioscience 6, a1-12, April 1, 2001]

CLUSTERING AMINO ACID CONTENTS OF PROTEIN DOMAINS: BIOCHEMICAL FUNCTIONS OF PROTEINS AND IMPLICATIONS FOR ORIGIN OF BIOLOGICAL MACROMOLECULES

Ivan Y. Torshin

Laboratory Of Chemical Kinetics And Catalysis, Chair Of Physical Chemistry, Chem. Dept. of Moscow State University, Moscow, 119899, Russia and Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, PA 19107, USA.

TABLE OF CONTENTS

1. ABSTRACT

Structural classes of protein domains correlate with their amino acid compositions. Several successful algorithms (that use only amino acid composition) have been elaborated for the prediction of structural class or potential biochemical significance. This work deals with dynamic classification (clustering) of the domains on the basis of their amino acid composition. Amino acid contents of domains from a non-redundant PDB set were clustered in 20-dimensional space of amino acid contents. Despite the variations of an empirical parameter and non-redundancy of the set, only one large cluster (tens-hundreds of proteins) surrounded by hundreds of small clusters (1-5 proteins), was identified. The core of the largest cluster contains at least 64% DNA (nucleotide)-interacting protein domains from various sources. About 90% of the proteins of the core are intracellular proteins. 83% of the DNA/nucleotide interacting domains in the core belong to the mixed alpha-beta folds (a+b, a/b), 14% are all-alpha (mostly helices) and all-beta (mostly beta-strands) proteins. At the same time, when core domains that belong to one organism (*E.coli*) are considered, over 80% of them prove to be DNA/nucleotide interacting proteins. The core is compact: amino acid contents of domains from the core lie in relatively narrow and specific ranges. The core also contains several Fe-S cluster-binding domains, amino acid contents of the core overlap with ferredoxin and CO-dehydrogenase clusters, the oldest known proteins. As Fe-S clusters are thought to be the first biocatalysts, the results are discussed in relation to contemporary experiments and models dealing with the origin of biological macromolecules. The origin of most primordial proteins is considered here to be a result of co-adsorption of nucleotides and amino acids on specific clays, followed by en-block polymerization of the adsorbed mixtures of amino acids.

2. INTRODUCTION

In most cases, a one-domain protein is classified into one of the following structural classes: alpha, beta, alpha+beta, and alpha/beta. The structural class of a protein correlates well with its amino acid composition (1, 2). When amino acid composition of each protein of a small set was represented as a dot in a 20-dimensional space, the dots corresponding to the alpha, beta, and alpha/beta types were found to be located in different regions in this composition space, while proteins of the alpha + beta type were widely scattered in the space (3).

The most significant problem in the elaboration of a reliable scheme to predict structural class using only amino acid composition was choice of the similarity measure between any two proteins. Similarity measure is

usually represented as a distance in a multidimensional (20D) space. Such predictions may have definite practical importance, as, for example, in the characterization of proteins with no known sequences. However, data of the next level (amino acid sequence) allow prediction not only of the structural class or "fold", but also of the whole structure, using new methods in structural genomics (4). Anyway, the plain application of Euclidean distance measures for any two proteins does not seem to be highly successful in predictions. The most successful similarity measures (2) require definite "training sets" which may be selected in various ways. In addition, the prediction is based on predefined, inflexible schemes of dividing proteins in a fixed number of classes and then specifying a distance measure for each class, as in (2).

Thus, two sides of the paradox may be stated thus: "one cannot cluster the proteins without distance measure", yet "distance measures may be available only for specific groups of (already) clustered proteins". To resolve the paradox, let us consider 20D space as a kind of "phase diagram". Phase diagram models are used in physical chemistry for detailed description of phase-phase transitions. Such a diagram will comprise the section of a larger diagram, a graphic representation drawn for particular pressures, temperatures and pH. Using this analogy, amino acids will be "components", structural classes (or other general functional characteristic of proteins) may be termed as "phases", and contacts between the "phases" are "phase-phase contacts". In this context, distance measures, specific for each class (as it was proposed in (2)) do make sense, as they represent a function for describing properties of a definite phase. In other words, these specific measures could be applied to measure "distance" only inside each "phase", a measure denied by “phase” boundaries. Plain Euclidean measure also could not be applied without significant errors as each phase has specific properties.

The general topology of any phase diagram is described by the phase rule, which states that the number of phases plus number of physical variables is equal to number of components plus 2. As we consider the diagram as a specific section, the number of physical variables is 0: thus there may up to 22 "phases" or "protein structural classes" related to amino acid composition and not merely four or five. However, in physical chemistry, a single phase is defined as consisting of matter that is continuously uniform in chemical and physical properties. As definitions of “continuity” and "uniformity" of physical and chemical properties do depend on resolution, the number of phases will also depend on resolution. Thus, there may be much more than 22 “phases”, in the scope of our analogy. Another aspect to consider is what would be common between the proteins in one such "phase" (if proteins form some phases in a continuous range of contents).

To investigate these questions in detail, it is necessary to map a large database of amino acid contents into the 20D space. Although this kind of work has already been attempted (3), only 135 selected proteins were used for the mapping. In addition, the goal of that work (3) was prediction of structural class. Purposes of this work are: 1. Study of the functional significance of specific amino acid contents and 2. Preparation of a “crude” map of the whole “space”, rather than elaborating a prediction stratagem. The set of proteins used here contains more than 3,000 protein domains from a non-redundant dataset.

In order to cluster proteins a rule called "the microcontinuity principle" is proposed here. Although an Euclidean measure was found to be ineffective when used at large scale, such as measuring distance between the phases or between distant points inside one phase, it still could be used to cluster neighboring dots (proteins) into the phases. The 20D space is quantized, that is, divided into the number of cells. The microcontinuity principle suggests that any two adjacent and non-empty cells (each containing one or several proteins) may correspond to one “phase”.

Exploring the phase diagram analogy further, we could indeed consider that proteins had been formed from mixtures of amino acids at certain prebiotic conditions. This has been proved experimentally, at least for specific amino acids and mixtures (for example, see (5)). Thus, this study could have implications for an analysis of the origin of biological macromolecules and consequently, the origin of cellular life.

3. METHODS

3.1. Databases

Spatial structures were obtained from the PDB (6), and domain definitions were taken from the SCOP database (7). Several databases were used for obtaining reliable biochemical information: SWISSPROT (8), LIGAND (9) and BRENDA (online version of "Enzyme Handbook" (10), available at http://www.brenda.uni-koeln.de/). A list of non-redundant PDB entries (which is a part of the VAST database (11)) was used to obtain non-redundant sets of SCOP domains.

3.2. Clustering procedure

Amino acid contents of each protein were represented by a string of float values (a dot or a vector in 20D space). Contents of each protein were mapped into an associative array. Coordinate axes (each representing contents of an amino acid) were divided into a specified number of 1D-cells. Before mapping the set the coordinate axes were scaled by calculating maximal occurrences for each amino acid using all proteins of the dataset. After scaling, the amino acid composition of each protein was converted into the array coordinates to find coordinates of the correspondent 20D-cell. Each 20D-cell of the array contained a list of protein identifiers.

Dynamic classification (clustering) was based on the microcontinuity principle. Although a Euclidean distance measure could not be applied on a large scale because the space is non-isotropic, it may be applied locally to find the cells/dots that are adjacent to one another. The maximal distance between two adjacent cells was defined as the size of the diagonal of the smallest hypercube, that equals to the square root of 20 (for 20 amino acids). For each pair of occupied cells the Euclidean distance was calculated and if the distance was less than or equal to the maximal distance, the two cells were marked as adjacent.

Table 1. Dependence of the number of clusters and size of the largest cluster on the axis scaling parameter.

Run	Scaling parameter	Number of clusters	Size of the largest cluster
1	10	200	1600
2	14	270	668
3	15	292	360
3a	16	288	259
4	17	278	133
5	18	271	64
6	19	252	5
7	20	233	5
8	21	234	3
9	22	219	1

Total number of the domain entries used: 3164. "scaling", value (number of 1D cells per axis); "Nc",; "largest", size of the largest cluster.

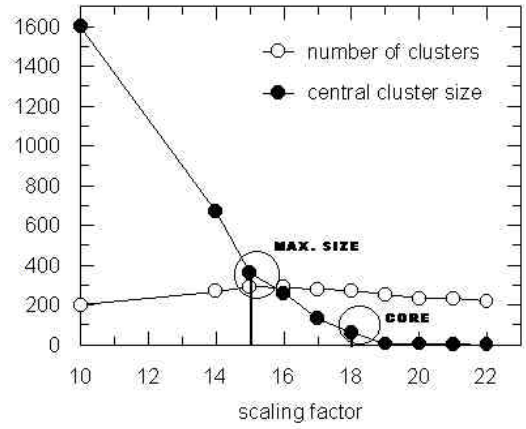

Figure 1. Number of clusters and size of the largest cluster depending on the scaling parameter. Minimal size of the central cluster ("core") and maximal size of the central cluster are marked.

As the distance measured was between the cells rather than the proteins, integer coordinates were used. For clustering, a tree-search procedure was applied after finding all adjacent cells.

The only empirical parameter in this approach was the number of cells per axis. This number was the same for all axes. The parameter, which is an analog of a "resolution", was varied and clustering was performed at different “resolutions”. Comparison of the results obtained at different resolutions has allowed the identification of the core of the largest cluster. All the programs used were developed by the author and were written in object-oriented Pascal.

4. RESULTS

4.1. Clustering procedure: the central cluster and its core

The results of clustering using different values of the scaling parameter are presented in table 1 and figure 1. Independently of the value of the scaling parameter, in all runs only one “largest” cluster was found. It was surrounded by hundreds of very small clusters (table 1). The total number of clusters grows until the 3rd run, and falls thereafter. This is because the minimal "cluster" size is 2 cells, and separate cells corresponding mostly to one protein were excluded. Therefore, at small values of the scaling parameter ("low resolution") these excluded cells tend to aggregate into larger clusters, thus providing very narrow "bridges" between the smaller clusters. As resolution grows, the number of abundantly aggregated cells is diminished and the number of clusters grows. After reaching a maximum at the 3rd run the number of clusters then drops as the clusters "disaggregate" at higher resolutions. Thus, the maximal size of the largest or "central" cluster size is that in the 3rd run, i.e., approximately 360 protein domains.

An important feature of the experiment is the acute (almost 12-fold) drop in the size of the largest cluster after the 5th run (figure 1). In the following runs all clusters are small. The central cluster is separated into small sub-clusters, each containing no more than 5 proteins. Most of the clusters contain only one protein or different variations of one protein. Thus, the smallest size of the central cluster was 64 domains. This collection of the domains was termed as "core" in the following text. Comparisons of the cluster contents from the 3rd through to the 5th run had shown that there are at least 75% of common entries in the central clusters of the runs 4 and 5, and 95% of common entries in the central clusters of the runs 3 and 5. Thus, the central cluster was 75%-95% conserved, independently on the value of the scaling parameter.

The whole picture was preserved, if instead of an 80% non-redundant set, a 100% non-redundant set of domains is used in clustering (all non-identical sequences were used). In this case also only one “largest” cluster was registered (the 2nd largest was a set of multiple lysozyme on-residue mutants, a feature of the PDB database). Contents of the core were almost the same, only some new proteins were added.

4.2. Amino acid contents of the central cluster and the core

The largest cluster is located in a specific region of amino acid contents (table 2). From the 5th to the 3rd run the boundaries are somewhat enwidened. Widths of the composition intervals however, are more or less similar for all amino acids. The cluster is placed in the region of

Table 2. Amino acid contents of the central cluster, the core and of the whole domain set (80% non-redundant set)

Amino acid	5.53 (core)	4.28	3.4	Whole set
GLY	0.04 0.09	0.03 0.11	0.03 0.14	0.00 0.23
ALA	0.04 0.12	0.03 0.13	0.02 0.18	0.00 0.28
VAL	0.05 0.11	0.04 0.10	0.03 0.12	0.00 0.21
LEU	0.05 0.13	0.05 0.14	0.04 0.14	0.00 0.25
ILE	0.03 0.09	0.03 0.09	0.01 0.12	0.00 0.19
SER	0.02 0.08	0.02 0.09	0.02 0.10	0.00 0.22
THR	0.03 0.09	0.02 0.08	0.02 0.10	0.00 0.21
ASP	0.04 0.09	0.03 0.10	0.02 0.10	0.00 0.21
ASN	0.02 0.07	0.02 0.08	0.01 0.08	0.00 0.14
GLU	0.04 0.10	0.02 0.11	0.02 0.12	0.00 0.26
GLN	0.01 0.06	0.01 0.08	0.01 0.08	0.00 0.16
LYS	0.03 0.11	0.03 0.11	0.00 0.12	0.00 0.24
HIS	0.01 0.05	0.00 0.05	0.00 0.06	0.00 0.26
ARG	0.01 0.08	0.01 0.08	0.01 0.09	0.00 0.21
PHE	0.02 0.06	0.02 0.07	0.00 0.07	0.00 0.12
TYR	0.02 0.06	0.01 0.08	0.01 0.08	0.00 0.12
TRP	0.00 0.03	0.00 0.03	0.00 0.04	0.00 0.09
CYS	0.00 0.05	0.00 0.05	0.00 0.07	0.00 0.35
MET	0.00 0.05	0.00 0.05	0.00 0.05	0.00 0.11
PRO	0.02 0.07	0.01 0.08	0.01 0.09	0.00 0.20

“5.53”, "4.28" and "3.4" are the largest clusters in the 5th, 4th and 3rd runs.

Table 3. Biosynthetic pathways of the amino acids

Pathway	Center	Branches
1	Asp:	Glu asn lys arg thr (ile, met)
2	asp:	gln (arg, pro)
3	asp:	ala ser gly
4	ala:	val leu
5	ala:	ser (cys, (tyr, phe, trp))
6	his	-

Amino acids adjacent in biosynthetic pathways are listed. Parentheses mark branching of the synthetic trees (according to (12)).

relatively low TRP, CYS and MET contents. The core is rather compact; the maximal volume of the core (product of the widths of all amino acids) is $4.18?10^{-26}$, while the whole set is $4.93\cdot10^{-15}$. When the dataset of non-identical sequences (instead of the 80% non-redundant dataset) was used, the composition boundaries of the largest cluster were only slightly enwidened. Below, individual amino acids are grouped into adjacent biosynthetic pathways (table 3) in order to consider tendencies in the amino acid contents (12). In the following text, pathway numbers are given in parentheses after amino acid labels. In the core set, ALA(3,4,5), LEU(4), and ARG(1,2) have the widest intervals, while ASP(1,2,3), GLN(2), HIS(6), PHE(5), TYR(5), CYS(5) and TRP(5) have the narrowest intervals. Thus, composition intervals of amino acid compositions of the 5th biosynthetic group are fixed.

The core boundaries extend from the 5th (core) to the 3rd run mainly as a result of the extension of GLY(3), ALA(3,4,5), ILE(1), GLU(1), LYS(1), ASP(1,2,3) contents. Contents of LEU(4), SER(3), THR(1), ASN(1), HIS(6) are increased less. This corroborates "fixed" composition intervals of the 5th group amino acids: the core does not “extend” through a content increase of "fixed" amino acids. It “extends” mostly by extension of composition intervals mostly of amino acids of the 1st group (and in some extent the 3rd), which are variable.

Contents of TRP(5), THR(1), MET(1), and ARG(1,2) remain almost the same while traversing from the core (5th run) to the whole cluster (3rd run). THR, MET are in the terminal branch, ARG is in the middle and ASP, GLU, ASN and LYS comprise the "stem" of the 1st group (pathway). The ARG interval is specific for the core/cluster. An attempt to analyze these data on the core contents is made in the Discussion section.

4.3. Structural classes, cellular locations and ligand-binding properties of the proteins of the core

Proteins comprising the core with domain data and biochemical data extracted from several databases are presented in the table 4. The SCOP database (7) was compiled using analysis of structural and some biochemical data. Although it is the most rationally constructed domain database, after visual inspection of the 64 "SCOP-domains" in the core cluster, only about 50 were found to be one-domain structures; the others were distinctly two-domain proteins. Domain definition is important as domains are probably independent folding units, and specific amino acid contents are characteristic for folding units with specific folds, rather than for the whole proteins (1).

In this work, the domains’ contents of the proteins of the core were analyzed visually and using a simple biochemical rule. If a protein has two geometrically distinct sub-domains it may be either a one-domain or a two-domain protein. If there are less than 2 covalent links between the geometrical sub-domains (which allows more independent folding), then the protein is considered as a two-domain protein. If more than 2, the protein is a one-domain protein with two sub-domains.

In the SCOP database the proteins are classified into 7 classes, the ones mentioned in the table 4 are classes 1-5 (percentages of proteins from the table 4 are give in parentheses): 1. All-alpha proteins (6%); 2. All-beta proteins (6%); 3. Alpha and beta proteins (a/b, 58%); mainly parallel beta sheets (beta-alpha-beta units); 4. Alpha and beta proteins (a+b, 25%); mainly anti-parallel beta sheets (segregated alpha and beta regions); 5."Multi-domain" proteins (alpha and beta, 4%, these are folds consisting of two or more domains/sub-domains of different classes). Thus, 83% of the core proteins fall into 3rd and 4th structural groups. The whole cluster (see runs 3 and 3a, table 1) contains about 72% of the (a/b) and (a+b) proteins. Nucleotide binding domains are frequently classified as being a/b or a+b.

Only 13% of the core proteins are periplasmic/extracellular proteins (1qaz, 1irp, 1azs, 3pva, 1aox, 1att), while most of the core proteins are intracellular. Data on cellular locations were extracted from SWISSPROT (8).

Data on ligand binding, extracted from several databases (see Methods) show that 64% of the core proteins interact with DNA/RNA/nucleotides; 15% with saccharides or polysaccharides; 9% are involved in protein-protein interactions and 12% have different and specific substrates of low molecular weight. More than half of the proteins of the last group have similar or close structural classes: 3.62

Table 4. Proteins of the core of the central cluster

SCOPid	ND	Ligand	SCOP class	Protein
1ft1b_	1	NADP	1.97.4.3.1	farnesyltransferase, beta {Rat}
1qaza_	1	scch	1.97.3.1.1	Alginate lyase {Sphingomonas}
1a59__	1	ADPi	1.98.1.1.4	CoA-citrate synthase {Antarctic bacterium}
1csh__	2m		1.98.1.1.1	Citrate synthase {Chicken}
1xbra_	1	DNA	2.2.5.1.8	transcription factor {African clawed frog}
1irp__	1	pp	2.40.1.2.3	IL-1 receptor antagonist {Human}
1rgs_1	1	cAMP	2.77.4.2.1	PKA regulatory subunit {Bovine}
7tima_	1	ATPi	3.1.1.1.3	Triosephosphate isomerase {Yeast}
1btc__	1	scch	3.1.7.2.1	beta-amylase {Soybean}
1qba_3	1	scch	3.1.7.6.1	Bacterial chitinase {Serratia}
1a4ma_	1	Nct	3.1.8.1.1	Adenosine deaminase {Mouse}
1adoa_	1	scch	3.1.9.1.5	Fructose-1,6-bisphosph. aldolase {Rabbit}
1dhpa_	1	sm	3.1.9.1.2	Dihydrodipicolinate synthase {E.coli}
1brla_	1	FMN	3.1.15.1.1	Bacterial luciferase {Vibrio harveyi}
1bsva_	1	NADP	3.2.1.2.3	GDP-fucose synthase {E.coli}
1qrra_	1	NAD	3.2.1.2.5	Sulfolipid biosynthesis SQD1 {Arabidopsis}
1dxy_1	1	NAD	3.2.1.4.3	D-2-hydroxyisocaproate dehydr. {Lactobacillus}
1gnd_1	2		3.3.1.3.1	G-nucleotide diss. Inhibitor {Bovine}
1nhp_1	1	NADH	3.3.1.5.7	NADH peroxidase {Streptococcus}
1b8ba_	1	Nct	3.6.1.3.1	ribonucleotide reductase-III {Bacteriophage T4}
1cm5a_	1	CoA	3.6.1.1.1	Pyruvate formate-lyase {E. coli}
1deaa_	1	scch	3.29.1.1.1	Glucosamine-6-phosphate deaminase {E.coli}
1dhs__	1	NAD	3.26.1.1.1	Deoxyhypusine synthase {Human}
1deka_	2m		3.31.1.1.3	Deoxynucleoside kinase {Bacteriophage T4}
1ftn__	1	GTP	3.31.1.7.8	RhoA, small gtpase family {Human}
1aoxa_	1	pp	3.57.1.1.5	Integrin a2-b1 {Human}
1c3ea_	1	Nct	3.60.1.1.1	gly-amide nucleotide transformylase {E.coli}
1bw0b_	1	sm	3.62.1.1.9	Tyrosine aminotransferase {Trypanosoma}
1cs1a_	1	sm	3.62.1.3.2	Cystathionine g-synthase {E.coli}
1qgna_	2		3.62.1.3.4	MalY gene expression reg {tobacco plant}
1bj4a_	2		3.62.1.4.8	Serine hydroxymethyltransferase {Human}
1bjna_	2m		3.62.1.4.6	Phosphoserine aminotransferase {E.coli}
1akn__	1	sm	3.64.1.1.4	Cholesterol esterase {Bovine}
1thtb_	1	sm	3.64.1.11.1	Myristoyl thioesterase {Vibrio harveyi}
2masa_	1	Nct	3.65.1.1.1	nucleoside N-ribohydrolase {Crithidia}
2pgi__	2m		3.74.1.2.2	Phosphoglucose isomerase {B.stear}
1fdo_2	1	NAD	3.75.1.1.3	Formate dehydrogenase H {E.coli}
1mioa_	2m		3.81.1.1.1	Nitrogenase Fe-Mo protein {Clostridium}
1ad3a_	2m	NADP	3.76.1.1.1	Aldehyde dehydrogenase {Rat}
8gpb__	2m	ATPi	3.82.1.2.1	Glycogen phosphorylase {Rabbit}
1dppa_	2m		3.89.1.1.10	Dipeptide-binding protein {E.coli}
1gr2a_	2m		3.89.1.1.16	Glu-receptor binding core {Rat}
1gcb__	2		4.3.1.1.8	Bleomycin hydrolase {Yeast}
1fumb_	2		4.13.6.2.5	Fe-S fumarate reductase {E.coli}
1gfla_	1	pp	4.20.1.1.1	Green fluorescent protein {Jellyfish}
1c8za_	1	DNA	4.21.1.1.1	Transcriptional factor tubby {Mouse}
1azsb_	1	pp	4.48.24.1.3	Adenylyl cyclase IIC1, domain C2a {Rat}
1eps__	2		4.55.2.2.3	synthase of a shikimate-derivative {E.coli}
1bkca_	1	pp	4.76.1.8.1	TNF-alpha converting enzyme {Human}
1tis__	1	UMP	4.96.1.1.4	Thymidylate synthase {Bacteriophage T4}
1bgya_	2		4.111.1.1.1	Core 1 subunit {Bovine}
1dik_3	1	ATP	4.121.1.5.1	Pyruvate phosphate dikinase {E.coli}
1ako__	1	DNA	4.130.1.1.1	DNA-repair exonuclease III {E.coli}
1ryp2_	1	ATP	4.132.1.4.2	Proteasome beta subunit {Yeast}
1rypg_	1	ATP	4.132.1.4.4	Proteasome alpha subunit {Yeast}
3pvaa_	1	sm	4.132.1.3.1	Penicillin V acylase {B.sphaericus}
1msk__	1	FAD	4.151.1.1.1	Methionine synthase {E.coli}
1atta_	1	pp	5.1.1.1.5	Antithrombin {Bovine}
1ceza_	1	RNA	5.8.1.3.1	T7 RNA polymerase {Bacteriophage T7}
1cc1s_	2		5.16.1.1.3	Fe-Ni hydrogenase {Desulfomicrobium}
1amub_	2		5.20.1.1.2	gramicidin synthetase 1 {B.brevis}
1ba3__	2		5.20.1.1.1	Luciferase {Firefly}

Protein domains that contain lost sequences and same proteins with observably different amino acid composition were excluded. ND- number of domains (after visual inspection); "2m" in this field stands for a two-domain protein with a middle domain. Ligand data are given only for one-domain proteins. "pp": protein-protein interactions; "scch": saccharide or polysaccharide; "sm": "specific metabolite"; "Nct": nucleotide or nucleoside; "ATPi": ATP as inhibitor.

and 3.64 (table 4). DNA/RNA/nucleotide binding proteins are present in each structural class, but mostly in classes 3 and 4.

Proteins of the core were extracted from different source organisms. When proteins from only one organism are taken (*E.coli* is the most frequent source in the table 4) then up to 80% of the core proteins are DNA/nucleotide interacting proteins. Thus, the main conclusion is that amino acid contents of most of the core proteins are not "random", they may be adapted to form specific protein templates for nucleotide recognition. The proteins of the "core" also seem to be "adapted" to their cellular location.

4.4. Proteins of the core that contain Fe-S clusters

Ferredoxins are small electron transporting proteins that have high proportion of smaller and dynamically stable amino acids and normally contain two or more, iron-sulfur clusters (a ferredoxin from *Thermotoga maritima* is unusual in that it contains a single cluster (13). Ferredoxins may have been the earliest biological catalysts prior to NADPH (14,15,16). CO-dhase (carbon monoxide (CO) dehydrogenase) is an iron-sulfur protein that, along with ferredoxin, thought to be one of the first proteins in primordial energy metabolism (17, 18). Thus, proteins with Fe-S clusters were, probably, vital in early cellular processes.

Four Fe-S containing proteins were found to be inside or near the core. Although they were classified as 1-domain in SCOP, some of them were identified as being two-domain after the visual analysis. Fumarate reductase (1fumb_, table 4) has 2 domains (A1-M106; T107-R243) and it binds 3 Fe-S clusters. Nickel-iron hydrogenase (1cc1s_) has 2 domains or sub-domains (K6-F202; F203-E283) and the whole molecule binds 3 Fe-S clusters. The smaller 80-residue sub-domain is significantly distorted, so it may, or may not, be an independent "folding unit". Formate dehydrogenase H (1fdo_2) has one domain that binds one Fe-S cluster. This is a one domain-protein as it has 3 covalent links between the sub-domains and not 2 as in most of the domains. Class III anaerobic ribonucleotide triphosphate reductase (1b8ba_) is a one-domain iron-sulfur protein.

As some of the two-domain proteins were clustered as one-domain proteins, the amino acid composition of each separate domain was compared with that of the core. The results are presented in the table 5. Amino acid contents of formate dehydrogenase H (1fdo_2) and ribonucleotide reductase (1b8ba_) are within the core boundaries. Although sub-domains of 1fum_b and 1cc1_s differ somewhat in their GLN contents, their amino acid contents are adjacent to the core cluster contents. Positions of ferredoxin and CO-dhase clusters in relation to the core were analyzed (table 5). The ferredoxin cluster is adjacent to the core cluster: for almost all amino acids, content ranges of the core overlap with the ferredoxin cluster. ASP and CYS contents of the ferredoxin cluster are slightly higher. CO-dhase contents overlap with the cluster's boundaries; CYS also has slightly higher contents.

Thus, there are at least five Fe-S binding domains with contents in the range of the core set. Amino acid contents of ferredoxin and CO-dhase clusters, thought to be one of the first proteins, overlap with the core boundaries. Thus they have amino acid contents similar to those of the core proteins.

5. DISCUSSION

5.1. Protein biochemistry and clustering of amino acid contents

There is no basis to claim that 20D space of amino acid contents is an orthogonal and an isotropic space (the one in which "euclidian" or "pythagorean" distance measure could be used). Moreover, there are definite physico-chemical reasons to assume this space to be non-orthogonal and non-isotropic: namely, the presence of "phase contacts" or "phase boundaries". Therefore, a simple distance rule like Euclidean distance could not be used to measure the distance between any two points in such space adequately (as one point may be in the region of one phase and the other in a region of some other). Using distance measurement procedures like the Mahalanobis measure (2,19), although it does produce rather remarkable results for a set of proteins, requires some "training set" of the proteins. The selection of an adequate training set or database represents a separate problem: predictions seem to depend on a given selection (20). The proposed here classification scheme, based on the microcontinuity principle, which assumes "micro-isotropicity" of the space.

To ensure that there are continuous clusters and that the microcontinuity principle could be applied, it was necessary to take as many proteins as possible. As results show, the outcome (one giant central cluster with a relatively small core and hundreds of micro clusters) is independent on the used dataset (80% and 100% non-redundant datasets were used). Thus, in terms of our "phase diagram" analogy, there is one major "phase" (a central cluster that has more or less continuous contents) with varying amino acid contents and hundreds of other specific "phases" with more or less fixed amino acid contents. There are at least two main problems related to the task of dynamic classification of amino acid contents: domain (or folding unit) definition and source organisms of the proteins.

The domain definition has already been mentioned in the Results section. There are no completely automated procedures that would allow identification of domains as independent folding subunits for any given protein structure. Separate domains are most probably independent folding units. Thus, defining sequence boundaries of a domain requires at least visual analysis, while biochemical studies on folding were made only for some specific small proteins.

Using proteins from different sources for clustering also may lead to additional displacements of data. Proteins from different organisms have survived multiple changes corresponding to adaptation/genetic selection on a biological level. Amino acid contents of proteins from an organism strongly depend on the environment: thermophiles, for example, have higher contents of polar amino acids, cysteine contents are related

Table 5. Amino acid composition boundaries of the central core, of a ferredoxin cluster and Fe-S domains from the proteins of the core

Amino acid	5.53	ferredoxin	1fumb1	1fumb2	1cc1s1	1cc1s2	1fdo_2	1b8ba	1qj2.fe_s
GLY	0.04 0.09	0.03 0.07+	0.04+	0.07+	0.08+	0.09+	0.08+	0.07+	0.09 0.12+
ALA	0.04 0.12	0.06 0.10+	0.09+	0.10+	0.10+	0.07+	0.10+	0.06+	0.09 0.10+
VAL	0.05 0.11	0.07 0.08+	0.08+	0.04	0.09+	0.05+	0.07+	0.06+	0.03 0.06+
LEU	0.05 0.13	0.07 0.10+	0.09+	0.05+	0.10+	0.02	0.06+	0.08+	0.07 0.08+
ILE	0.03 0.08	0.04 0.07+	0.04+	0.07+	0.05+	0.04+	0.06+	0.07+	0.07 0.09+
SER	0.02 0.08	0.01 0.07+	0.05+	0.07+	0.05+	0.04+	0.05+	0.06+	0.04 0.05+
THR	0.03 0.09	0.02 0.08+	0.05+	0.05+	0.05+	0.04+	0.06+	0.06+	0.05 0.10+
ASP	0.04 0.08	0.09 0.12	0.08+	0.04+	0.05+	0.07+	0.05+	0.06+	0.01 0.04+
ASN	0.02 0.07	0.01 0.05+	0.06+	0.05+	0.03+	0.09	0.05+	0.06+	0.04 0.05+
GLU	0.04 0.10	0.08 0.14+	0.06+	0.05+	0.08+	0.09+	0.06+	0.06+	0.05 0.06+
GLN	0.02 0.06	0.03 0.05+	0.00	0.07	0.01	0.00	0.04+	0.03+	0.04 0.06+
LYS	0.03 0.09	0.04 0.06+	0.06+	0.06+	0.05+	0.07+	0.05+	0.08+	0.03 0.03+
HIS	0.01 0.05	0.00 0.02+	0.01+	0.04+	0.05+	0.01+	0.02+	0.03+	0.03 0.06+
ARG	0.01 0.08	0.01 0.03+	0.05+	0.04+	0.03+	0.02+	0.05+	0.04+	0.03 0.06+
PHE	0.02 0.06	0.01 0.05+	0.03+	0.04+	0.04+	0.06+	0.04+	0.04+	0.03 0.04+
TYR	0.02 0.06	0.01 0.06+	0.06+	0.04+	0.02+	0.05+	0.04+	0.04+	0.00 0.04+
TRP	0.00 0.03	0.00 0.02+	0.01+	0.01+	0.02+	0.02+	0.02+	0.01+	0.00 0.00+
CYS	0.00 0.04	0.03 0.09+	0.04+	0.05	0.03+	0.09	0.02+	0.02+	0.05 0.05
MET	0.00 0.05	0.00 0.01+	0.06	0.01+	0.03+	0.00+	0.03+	0.03+	0.04 0.05+
PRO	0.02 0.07	0.03 0.08+	0.06+	0.06+	0.07+	0.07+	0.04+	0.04+	0.03 0.06+

If contents of an amino acid are within the range of the core contents, they are marked with "+". 1qj2.fe_s are contents of the Fe-S domains from a CO-dhase (CO-dehydrogenase, PDB 1qj2).

to organism's complexity. In this work, only 83% of the core proteins have mixed alpha, beta folds, while all core proteins from *E.coli* have these specific folds. 64% of the core proteins (all organisms) are involved in nucleotide binding, while 80% of *E.coli* (one organism) proteins have the same function. Therefore, in order for procedure to be biologically complete, it would be necessary to map into the array domains of a small genome: for example, archaea or protea. In this case, however, there would be almost no available structural data to identify domains, as the genome sequences have been elucidated only recently, not to mention spatial structures.

5.2. Amino acid composition, structural class and function of proteins

Amino acid composition itself, even in the absence of sequence data, is a significant factor in determining structural class and therefore, to some extent, general function(s) of a protein. Exhaustive enumerations of all conformations of all sequences using simple lattice models consisting of two types (hydrophobic and polar) residues were made for specific amino acid contents (21). The proteins belonging to the four distinct folding classes were shown to have significant differences in their distributions of non-bonded contacts. These differences provide a physico-chemical basis for the correlation between structural class and amino acid composition (21).

"Structural class" is a specific category which may or may not be related to function(s) of the protein. A study of correlation, for example, between enzymatic class and structural class, has shown no strong preferences (22). For several common ligands a distinct preference of a certain fold was found: hem- all-alpha domains; DNA-binding and nucleotides- all-alpha and mixed a,b domains. The present work shows that similar function (such as nucleotide binding) of different proteins may be closely related to the similar amino acid contents.

Amino acid contents may define not only the specific function of a protein but also its cellular location. Correlation analysis of amino acid composition and the cellular locations of proteins discriminates among the following five protein classes: integral membrane proteins, anchored membrane proteins, extracellular proteins, intracellular proteins and nuclear (for eukaryotes) proteins (23). The bulk (87%) of the proteins in the core of the central cluster that has specific amino acid contents are intracellular (both in prokaryotes and eukaryotes) and many of them are nuclear (in eukaryotes) proteins.

Proteins of the core are characterized by relatively low TRP, CYS and MET contents (in comparison to the whole set). Low contents of TRP and MET are also characteristic of the ferredoxin cluster and CO-dhase (table 5), the oldest iron-sulfur proteins. At the same time, CYS contents of these oldest proteins are higher than the CYS contents of the cluster, although both MET and CYS are kinetically unstable amino acids. The following section dealing with Fe-S proteins presents some reasons for the higher CYS contents.

Among the core proteins, contents of ASP, GLU, ASN and LYS, comprising the "stem" of the 1st pathway (table 3) are the most variable. ASP, ALA, GLY of the 3rd group are variable. PHE(5), TYR(5), TRP(5) CYS(5), consecutive stages comprising the top, or "crown" of the 5th biosynthetic group, are observably less variable. ARG(1,2), THR(1) and MET(1) that form the "crown" of the 1st pathway and are almost invariable. Almost all the non-variable or less variable amino acids have low percentages in ferredoxins (primarily, ARG, MET, PHE, TRP, CYS and, to some extent, TYR). Non-variability of an amino acid content may correspond to the low specific concentration of the amino acid available prior to polymerization or simply to the absence of this particular

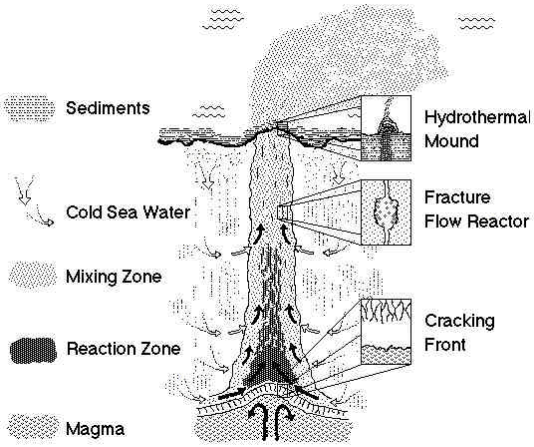

Figure 2. A submarine hot spring reactor. Seawater descends into cold permeable rock from the sea floor and is drawn into the hydrothermal system by the siphoning effect of the rising hot fluid. The thermal energy is supplied by the advance of the cracking front into the magma body, the top of the magma body (~1200 C) is 1-5 km beneath the sea floor. The hydrothermal vent (hot spring) hypothesis (23) proposes that: (1) At the cracking front sea water is heated rapidly to ~600 C, carbon is extracted from the rock, reactions with ferrous iron in the rock produce a reduced fluid containing significant quantities of methane, hydrogen, carbon dioxide and ammonia. Sea water is converted to fluid strongly enriched in silica and transition metals, and significantly enriched in calcium and potassium; (2) High-energy organic monomers are synthesized in the fluid at or in the vicinity of the cracking front; (3) The flow reactor solves the serious problem of short-lived intermediates (hydrogen cyanide, sugars, etc) in pre-biotic synthesis; (4) Fractures in the upper parts of the hot spring flow reactors can extract and accumulate organic matter, thus achieving high concentrations of the pre-biotic reagents. Synthesis of organic polymers (proteins and nucleic acids) is most plausible here; (5) The catalytic surfaces of clay minerals can provide initial information to organize organic components in the fluids. (courtesy of J. Corliss)

amino acid in the first formed (primordial) proteins. The second assumption has greater weight as these invariable and less variable amino acids are also, for the most part, dynamically unstable.

Thus, analysis of amino acid contents of the core cluster and comparisons with amino acid composition of the oldest proteins shows that amino acids from the "stem" of the 3rd and 1st biosynthetic pathways are likely to have available at an early stage of evolution, whereas the "crowns" of the 1st and 5th pathways may be a later addition.

This conclusion is consistent with models of the chemistries of hydrothermal solutions that exhaled from the Earth's crust about 4 billion years ago. Metal sulfides and oxides in the crust would not only have buffered hot water circulating in the crust, they may also have acted as catalysts in the synthesis of amino acids from atmospheric carbon dioxide and hydrothermal hydrogen and ammonia and minor cyanide (24, 25). Hydrothermal experiments in which millimolar quantities of KCN, NH_4Cl, H_2CO and NaHS were reacted with CO_2 and H_2 in the presence of pyrite (FeS_2), pyrrhotite (Fe_7S_8), magnetite (Fe_3O_4) and rutile (TiO_2) at 150°C, produced observable quantities of glycine, minor aspartic acid and alanine and trace serine (26, 27). All four of these amino acids are in the 3rd biosynthetic pathway and are variable amino acids of the core. Moreover, these are just the amino acids previously assumed to comprise the repeating sequence from which a protoferredoxin may have been emerged (14).

5.3. Origin of cellular life in hydrothermal vents and Fe-S proteins

The recent discovery of 3,235-million-year-old pyritic microfilaments (28) shows that the ancient microorganisms were probably thermophilic and chemolithotrophic prokaryotes that inhabited hydrothermal vents of the ocean floor. Such environments may have been hosts for the first living systems on the Earth and this is highly consistent with models of a thermophilic origin of life in hydrothermal vents (hot springs) (18,28,29). A scheme of a hydrothermal feeder and mound, according to (30), is presented in the figure 2 (an animation of a present day "smoker" is available in "Into The Abyss", online presentation by NOVA TV, at URL: http://www.pbs.org/wgbh/nova/abyss/life/extremes.html). Some of the simplest and genetically most primitive bacteria use sulfur in their metabolic processes. Many of such bacteria live at hot spring sites. As now, reduced iron and sulfur would have been common constituents of the primordial hydrothermal vents. These two elements exhibit variable valences and may have been vital to the emergence of cellular life (18).

A universal ancestral metabolic complex (UAMC) could have being employed in the channeling of metabolites (31). The intermediate molecules in the first metabolic pathways may have being transferred directly, without diffusion, between catalysts localized at specific sites of a mineral surface. The nano-crystalline mackinawite (figure 3a), which is very similar in structure to clay minerals, would form the basic structure for the UAMC. In this environment mackinawite ($Fe(Ni,Co)_{1+x}S$) is easily oxidised to greigite (32). Greigite ($[Fe_4S_4][SFeS]_2$) and the nickel-bearing homologue violarit ($[Fe_2NI_2S_4][SFeS]_2$) (25) (figure 3b) catalyses the reversible oxidation of CO to CO_2 as well as condensing carbon monoxide, a methyl group and Co-A, to form acetyl Co-A (17). Amino acid contents of the core cluster that contains several Fe-S binding proteins, but comprise mostly DNA-binding proteins, overlap with the content boundaries of the CO-dehydrogenase and ferredoxin clusters. Positively charged mineral particles from which Fe-S clusters (figure 3c) would have being formed may have acted as the first catalysts prior to NADs and other nucleotide cofactors (16). Because of their positive charge (e.g., as $[Fe_4S_4]^+$ (32)), the first proteins would be formed using Fe-S clusters as precipitation centers and, later, nucleotides may have substituted the Fe-S clusters

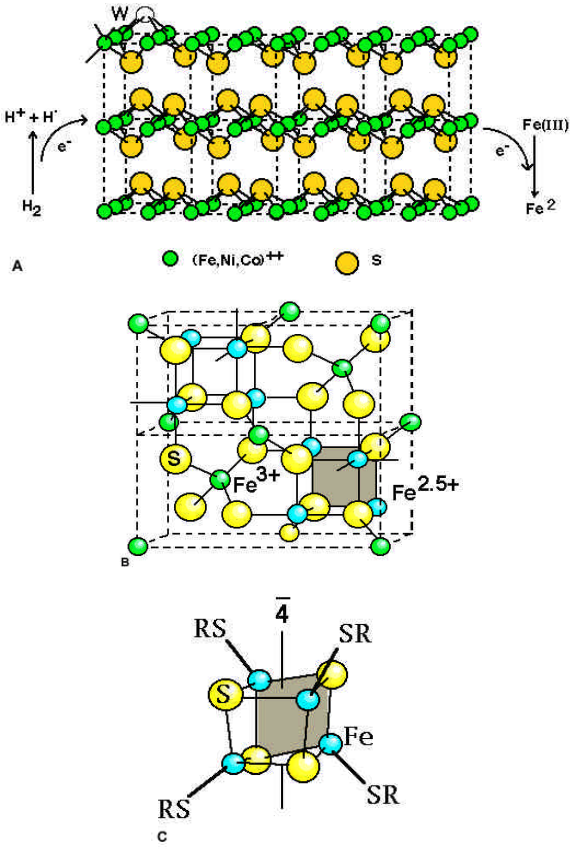

Figure 3. Iron-sulfur minerals that could be important for origin of cellular life in a hydrothermal vent (hot spring). a) Structure of mackinawite. Mackinawite could have provided the basic structure for a universal ancestral metabolic complex (UAMC). b) Structure of greigite. c) Fe-S cluster in ferredoxins. (b) and (c) show the similarity in structure of the Fe_4S_4 "thio-cubane" cluster in ferredoxins (c) with the $Fe_4^{2.5+}S_4$ "cubane" unit of greigite, Fe_3S_4 (b). Fe_4S_4 is variably distorted in ferredoxins depending on the oxidation state. The proto-ferredoxin would be intermediate in oxidation state between more oxidized greigite (b) and mackinawite (a), which has a layered structure (25). ((a),(b),(c) Courtesy of M. Russell and A. Hall).

in the protein moulds. Another possibility is discussed in the following section.

Ferredoxin and CO-dhase have higher CYS contents than the core proteins (table 5). Although cysteine is an unstable amino acid and thus seemingly could not be presented in the oldest proteins, most cysteines of ferredoxins from older organisms are involved in binding Fe-S clusters. Iron atoms of Fe-S clusters are bound exclusively by these cysteines. This suggests that the unstable cysteines could have been stabilized by interactions with Fe-S clusters.

5.4. Formation of protein-DNA/nucleotide interface

The DNA-protein interface is the central feature of all known cellular life. The important question is how this interface, characteristic for any cell, may be formed. Two ways of generation are considered here: either selection of "ready-made" proteins with specified amino-acid contents ("evolutionary" selection) or, *in situ*, spontaneous synthesis of the amino acid chains from co-adsorbed mixtures of nucleotides and amino acids (e.g. 33,34). The adsorbed mixtures would have specific amino acid composition contents, a result of specific interactions with nucleotides, minerals and between amino acids in the mixture. The first assumption is inevitably related with a "two-world" or dualistic model(s): some "RNA/DNA-world" and some "protein world". This dualistic model has certain advantages: RNA indeed may have formed a separate "world" and could self-reproduce. The main point to be held in view, however, is that outside of a modern cell nucleic acids molecules and proteins require different conditions for their "survival". Polynucleotide molecules are delicate and easily damaged and could not survive alone in the harsh environments at hydrothermal vents (with temperatures of at least 100°C). As experimental evidence is consistent with life having originated in hydrothermal vents or seepages (28), the RNA world(s) is thermodynamically improbable.

In the model of in-place synthesis of the amino acid chains from the co-adsorbed monomers, proposed here, mixtures of amino acids and "future ligands" of the “future protein molecules”-such as nucleotides- would adsorb on minerals and form complexes, which may appear similar to modern binding sites excised from the rest of a protein. The composition of such mixtures certainly depends on the positively charged nucleotides as well as properties of the mineral templates. Binding site geometry in modern proteins that could be constituted in a number of ways (22) corresponds here to different structuring of a mixture and/or to different mixtures co-adsorbed at different micro-conditions such as, for example, surface properties of the mineral. Thus, the short primordial oligonucleotides may have been synthesized from nucleotides along with protective protein shells and this certainly may contribute to DNA's stability at higher temperatures.

Expansion of the central cluster (depending on resolution) leads to inclusion of at least 10% of proteins from the whole domain set into the central cluster (table 1). This observation may correspond to many proteins having a single common origin. And this origin is interrelated with the formation of a nucleotide (and /or DNA)-protein interface. In other words, structural organization of “native”, “biological” proteins may be intrinsically linked to chemical properties of nucleotides.

5.5. En-block polymerization of the amino acids in amino acid/nucleotide mixtures co-adsorbed on clay and formation of non-random amino acid sequences

One of the possible chemical scenarios to form the first DNA-protein complexes is synthesis inside a microcavity of a mineral. This microcavity would have the size of a protein: nucleotide(s) could be co-adsorbed with mixtures of amino acids inside it. Some of the amino acids would interact with nucleotide, some with the walls of the cavity. Thus, folding of primordial proteins would not be

separated from their synthesis. The synthesis would be due to some kind of en-block polymerization of the mixtures. As there is a number of kinetically unstable amino acids, primordial “alphabet” of amino acids would not be as extensive as 20 amino acids (25).

There are several experiments that provide partial support for this co-adsorption model. Co-adsorption of the nucleotides and proteins may be regulated due to small fluctuations of pH such as occur in hot springs or seepages. These fluctuations could lead to cyclical adsorption-desorption and accumulation of nucleotides on the fracture walls of the hot springs (35). Montmorillonite, a type of clay likely to be present in the walls of hot springs, can act as a heterogeneous catalyst for aqueous polycondensation of amino acids and as an inhibitor of competitive hydrolysis (36). During experiments a discrete grouping of molecular weights was noted. Pulsation of pH may lead to simultaneous formation of peptides and polynucleotides during an oscillatory catalytic process on the clay (36).

Transition metals, abundant in submarine hot springs, have been shown to be effective catalysts of amino acid condensation. Proteinoid proto-shells, synthesized by wet-drying cycles and thought to be primordial cell membranes (5), have also been synthesized from amino acids in a seawater medium. The medium is modified by addition of transition metals and held at pH 5 and 105^{o}C for several weeks- at conditions highly plausible for hot springs (37).

Spontaneously formed amino acid sequences are not necessarily "random". The information supplied by the patterns of charge distribution on the surfaces of mineral lattice planes may induce higher hierarchical level(s) in the organization of the biopolymers (30). Even without minerals, thermal polymerization, for example of GLU, TYR and GLY at 180^{o}C has resulted in preferential synthesis of peptides having certain sequences. The effects of amino acid auto-ordering could be a response to stereo-specific interactions between the amino acids (38).

6. PERSPECTIVE

The structural class of a protein, it's general biochemistry (particularly DNA-binding properties) as well as cellular location, are intrinsically related to the protein's amino acid contents. Using the structural analogy of “phase diagram” as a kind of guide, protein domains from a large dataset were mapped into 20D space according to their amino acid contents. Analysis of this map's connectivity at different resolutions has shown that there is one largest (or “central”) cluster and swarms of very small clusters for proteins with specific functions/ligands. Most of the proteins in the core of the central cluster are involved in DNA/nucleotide binding and the cluster has specific boundaries of amino acid composition. The cluster core is adjacent to the clusters of ferredoxins and carbon monoxide dehydrogenase, proteins with the oldest ancestry. These two proteins are thought to be central to primordial energy metabolism (14,16,32).

Developing experimentally supported models of origin of cellular life is very important for understanding of cellular life. It is important as well for the search for potential cellular life outside the Earth and thus for the new field of astrobiology. Experimental pre-biotic synthesis suggests that all the basic components of living cells could be produced and assembled in primitive hot springs. The hot spring models account for the synthesis, concentration and accumulation of organic matter from dilute solutions. Steady-state flow reactors just beneath hot springs (30) or seepages (18) allow participation of very short-living intermediates in extensive chains of homogenous and heterogeneous catalysis and thus are chemically plausible. The hot spring model also takes into account what the Earth was like over four thousand million years ago. According to geological data (18), it was a hot and humid place, with only few ephemeral volcanic islands in the acidic ocean covering the entire planet.

This paper seeks to provide one of a myriad of possible links between the chemical and bio-chemical worlds. Formation of a DNA-protein interface (characteristic of all known cellular life) is considered here as characterized by the spontaneous synthesis of amino acid chains from the mixtures of nucleotides and amino acids co-adsorbed on specific clays. Conditions in primordial hot springs would have encouraged such a synthesis ("en-block polymerization"). Amino acid polymerization would probably have been catalyzed by transition metals. Although there is some experimental evidence for the formation of nucleotides and polypeptides in such environment (35,36,37), polymerization of amino acid mixtures co-adsorbed with their ligands (nucleotides or Fe-S clusters) may require specific conditions such as a particular state of a mineral’s surface. This scheme also accounts for the higher stability of nucleotides/DNA's in such extreme environment as hot springs: the DNAs would be stabilized by the co-forming protein shell.

7. ACKNOWLEDGEMENTS

Author thanks Prof. M. Russell and Prof. A. Hall for providing figures of the mineral structures and the book chapter, Prof. J. Corliss for the scheme of a flow reactor in a hot spring and Prof. D. Caldwell for providing abstracts and summaries of his new unpublished papers.

8. REFERENCES

1. K. C. Chou & G. M. Maggiora: Domain structural class prediction. *Protein Eng* 11(7), 523-538 (1998)

2. K. C. Chou & C. T. Zhang: Prediction of protein structural classes. *Crit Rev Biochem Mol Biol* 30(4), 275-349 (1995)

3. H. Nakashima, K. Nishikawa & T. Ooi: The folding type of a protein is relevant to the amino acid composition. *J Biochem* (Tokyo) 99(1), 153-162 (1986)

4. L. A. Kelley, R. M. MacCallum & M. J. Sternberg: Enhanced genome annotation using structural profiles in the program 3D- PSSM. *J Mol Biol* 299(2), 499-520 (2000)

5. S. W. Fox & K. Dose: Molecular evolution and the origin of life, 2nd ed. Marcel Dekker, New York, 1-100 (1977)

6. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov & P. E. Bourne: The Protein Data Bank. *Nucleic Acids Res* 28(1), 235-242 (2000)

7. A. G. Murzin, S. E. Brenner, T. Hubbard & C. Chothia: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247(3), 536-540 (1995)

8. A. Bairoch & R. Apweiler: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(1), 45-48 (2000)

9. S. Goto, T. Nishioka & M. Kanehisa: LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res* 28(3), 380-382 (2000)

10. D. Schomburg, M. Saltzmann & D. Stephan: (editors) Enzyme Handbook. Springer-Verlag 1998-2000

11. J-R. Gibrat, T. Madej & S. H. Bryant: Surprising similarities in structure comparison. *Curr Opinion Struct Biol* 6(0), 377-385 (1996)

12. E. Szathmaary: The origin of the genetic code: amino acids as cofactors in the RNA world. *Trends in genetics* 15(6), 223-229 (1999)

13. B. Darimont & R. Sterner: Sequence, assembly and evolution of a primordial ferrredoxin from *Thermotoga maritima. EMBO J*,, 13, 1772-1781 (1994).

14. R.V. Eck & M.O. Dayhoff: Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science*, 152, 363-366 (1966).

15. D.O. Hall, R. Cammack & K.K. Rao: Role for ferredoxins in the origin of life and biological evolution. *Nature*, 233, 136-138 (1971).

16. R.M. Daniel & M.J. Danson: Did primitive microorganisms use nonheme iron proteins in place of NAD/P? *J Mol Evolution*, 40, 559-563 (1995).

17. P.A. Lindahl, E. Münk & S.W. Ragsdale: CO dehydrogenase from *Clostridium thermoaceticum.:* EPR and electrochemical studies in CO_2 and argon atmospheres. *J Biol Chem*, 265, 3873-3879 (1990).

18. M. J. Russell, A. J. Hall, A. G. Cairns-Smith & P. S. Braterman: Submarine hot springs and the origin of life. Correspondence. (for more details http://www.gla.ac.uk/Project/originoflife/qas.html) *Nature* 336(1), 117-118 (1988)

19. G. P. Zhou: An intriguing controversy over protein structural class prediction. *J Protein Chem* 17(8), 729-738 (1998)

20. F. Eisenhaber, C. Frommel & P. Argos: Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins* 25(2), 169-179 (1996)

21. I. Bahar, A. R. Atilgan, R. L. Jernigan & B. Erman: Understanding the recognition of protein structural classes by amino acid composition. *Proteins* 29(2), 172-185 (1997)

22. A. C. Martin, C. A. Orengo, E. G. Hutchinson, S. Jones, M. Karmirantzou, R. A. Laskowski, J. B. Mitchell, C. Taroni & J. M. Thornton: Protein folds and functions. *Structure* 6(7), 875-884 (1998)

23. J. Cedano, P. Aloy, J. A. Perez-Pons & E. Querol: Relation between amino acid composition and cellular location of proteins. *J Mol Biol* 266(3), 594-600 (1997)

24. M.D. Schulte & E.L. Shock: Thermodynamics of Strecker synthesis in hydrothermal systems. *Origins Life Evolution Biosph*, 25, 161-173 (1995).

25. J.P. Amend & E.L. Shock: Energetics of amino acid synthesis in hydrothermal ecosytems. *Science* 281, 1659-1662 (1998)

26. R. J. Hennet, N. G. Holm & M. H. Engel: Abiotic synthesis of amino acids under hydrothermal conditions and the origin of life: a perpetual phenomenon? *Naturwissenschaften* 79(3), 361-365 (1992)

27. A. L. Weber & S. L. Miller: Reasons for the occurrence of the twenty coded protein amino acids. *J Mol Evol* 17(5), 273-284 (1981)

28. B. Rasmussen: Filamentous microfossils in a 3,235-miilion-year-old volcanogenic massive sulphide deposite. *Nature* 405(8), 676-679 (2000)

29. J. B. Corliss, J. Dymond, L. I. Gordon, J. M. Edmond, R. P. von_Herzen, R. D. Ballard, K. Green, D. Williams, A. Bainbridge, K. Crane & T. H. van_Andel: Submarine thermal springs on the Galapagos Rift. *Science* 203, 1073-1083 (1979)

30. J. B. Corliss: Life is a strange attractor: the emergence of life in Archaean submarine hot springs (http://www.syslab.ceu.hu/~corliss/LSABook.pdf) online book, 67-94 (1990)

31. M. R. Edwards: Metabolite channeling in the origin of life. *J Theor Biol* 179(3), 313-322 (1996)

32. M. J. Russell, D. Daia & A. J. Hall: The emergence of life from FeS bubbles at alkaline hot springs in an acid ocean. In: Thermophiles: The keys to molecular evolution and the origin of life? Eds: Wiegel J, Adams MWW. Taylor and Francis, London and Philadelphia, 77-126 (1998)

33. A.R. Mellersh: A model for the prebiotic synthesis of peptides which throws light on the origin of the genetic code and the observed chirality of life. *Origins Life Evolution Biosph*, 23, 261-274 (1993)

34. R.D. Knight & L.F. Landweber: Rhyme of reason: RNA-arginine interactions and the genetic code. *Chemical Biol*, 5R, 215-220 (1998).

35. J. G. Lawless & N. Levi: The role of metal ions in chemical evolution: polymerization of alanine and glycine in a cation-exchanged clay environment. *J Mol Evol* 13(4), 281-286 (1979)

36. M. Paecht-Horowitz, J. Berger & A. Katchalsky: Prebiotic synthesis of polypeptides by heterogeneous polycondensation of amino-acid adenylates. *Nature* 228(272), 636-639 (1970)

37. H. Yanagawa & F. Egami: Formation of organized particles, marigranules and marisomers, from amino acids in a modified sea medium. *Biosystems* 12(3-4), 147-154 (1980)

38. J. Hartmann, M. C. Brand & K. Dose: Formation of specific amino acid sequences during thermal polymerization of amino acids. *Biosystems* 13(3), 141-147 (1981)

39. D. M. Karl, D. F. Bird, K. Bjorkman, T. Houlihan, R. Shackelford & L. Tupas: Microorganisms in the accreted ice of Lake Vostok, Antarctica. *Science* 286(5447), 2144-2147 (1999)

Keywords: Amino Acid Composition, Iron Sulphides; Protein Folding; Protein Synthesis; Substrate Specificity; Origin Of Cellular Life

Send correspondence to: Dr Ivan Torshin (NSC 444), Dept. of Biology, 402, Kell Hall, Georgia State University, 24 Peachtree Ctr Ave, Atlanta, GA 30303, Tel: (404) 651-0098, Fax: (404) 651-2509, E-mail: biotiy@hydra.cs.gsu.edu, tiy135@yahoo.com