ISSN 1870-4069

Physicochemical Compatible Motifs in Proteins Sequences

Jesús Alberto Correa-Morales[1], Eduardo M. Martin[2], Eunice Esther Ponce de León-Sentí[2], Rogelio Salinas-Gutiérrez[3]

[1] Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Maestría en Ciencias con opciones a la Computación, Mexico

[2] Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Departamento de Ciencias de la Computación, Mexico

[3] Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Departamento de Estadística, Mexico

al205239@edu.uaa.mx, {emmartin, eponce, rsalinas}@correo.uaa.mx

Abstract. The identification of tridimensional motifs, folding arrangements in protein structure has been a strenuous task. Recently *ab initio* protein structure can be elucidated by computational intelligence algorithms, though it still time consuming and poses new problems. Until enough protein structures have been solved, the identification and classification of tridimensional motifs will remain an arduous task. Therefore, it still important to rely on approaches that are independent of tridimensional information. A methodology that uses only the psychochemical properties of amino acid pairing is here described. This methodology work independently from structural data, maximizing the physicochemical compatibility among amino acid pairs. Therefore, it is far easier to implement, and results can be obtained in a shorter time. This approach returns sequence pairs with high compatibility, which might be part of a protein motif. These can aid in the identification and classification of protein subsequences.

Keywords: Optimization, motifs, proteins, genetic algorithm, psychochemical compatibility.

1 Introduction

Proteins are polymers of amino acids joined by a covalent bond known as peptide bond. Twenty different amino acids are commonly found in proteins, each amino acid has amino and a carboxyl group as well as a side chain name a radical.

The radical of each amino acid vary in structure, size, charge, and hydropathy which confer each amino acid of specific properties. A typical protein has between 150 and up

S	$C(S)$	W	
		w	c
	A:10020000000		
	C:00000000090		
	D:00020001000		
TTFEIPQNVCV	E:00050000000		
LPPAPQVNKCN	F:00100100000	TGVCV	66.67
LPPAPQVNKCN	G:00000005000	VTGVCV	61.11
PNTEVFTDVCV	I:00003200200	TEVVTGVCV	59.26
PNTEVVTGVCV	K:00000000200	VVTGVCV	58.73
PNTEVVTGICV	L:20000020000	PNTEVVTGVCV	58.59
PNTEVVTGICV	N:06000003002	EVVTGVCV	58.33
ANTDIILGVCV	P:42202100000	NTEVV	53.33
SNTDIILGVCV	Q:00000210000		
	S:10000000000		
	T:11600040000		
	V:00004320507		

Fig. 1. This an example of how the search for conserved sequences (**SCS**) algorithm works. From a multiple alignment of sequences S, a consensus matrix is acquired $C(S)$ (for reference only the amino acids that are present in S are shown here), from $C(S)$ a set of conserved sequences W of variable length are order by decreasing c value.

to a few thousand amino acids, the amino acidic composition of a protein will confer it of specific structural and functional properties. Interactions among amino acids from the same protein lead to a specific folding, hence structure and function.

To understand the complexity of protein structure and function, protein's tridimensional structure has been further subdivided, motifs being folding arrangement identifiable as substructures. Therefore, knowing the interprotein interactions among amino acids that conform motifs can provide important information for a given protein [2, 7, 9, 10].

Genetic algorithms have extensive applications in optimizing combinatorial problems [4], the strength of this algorithm reside in its exploration power and capability to escape from a local optimum. The most important element when trying to solve an optimization problem with a genetic algorithm is the modelling of the individual [11], the main points to be considered when dealing with this kind of algorithms is the design of the objective function, the solution's modelling, and the population's conformation. The use of genetic algorithms in the search for motifs has already been explored, though this has been made mainly in DNA sequences working with nucleic acids [5].

The aim of the present is to describe a methodology composed of three stages, which will allow the user to identify pairs of sequence that are mutually compatible in the protein sequence, which might be conforming a tridimensional structural motif.

Algorithm 1 Pseudocode for multiple point mimicry

Input: $selected\_population$
Output: $cross\_population$

$cross\_population$ is initialized as an empty list
for assign $i \leftarrow 1$ every second individual until P **do**
  if CP is greater than a random number between 0 to 100 **then**
    generates four different random numbers between 0 and Lv
    random numbers are saved in $points$
    order the numbers in increasing order
    assign $aux \leftarrow 1$
    while $individual_i$ is equal to $individual_{i+aux}$ **do**
      if aux is greater than or equal to P **then**
        $aux \leftarrow -1$
      increment the value of aux by one
    $child_1$ is made up of segments of $individual_i$, $individual_{i+aux}$, and $individual_i$ defined by the numbers in the even positions of $points$
    $child_2$ is made up of segments of $individual_{i+aux}$, $individual_i$, and $individual_{i+aux}$ defined by the numbers in the odd positions of $points$
    $child_1$ adds to $cross\_population$
    $child_2$ adds to $cross\_population$
return $cross\_population$

2 Methodology

To identify compatible motifs based on amino acid pairing physicochemical properties a three-phase approach is devised, each with a specific objective. The output will be a set of highly compatible pair of sequences an its position in a multiple sequence alignment, these pairs will consist of a highly conserved sequence and a sequence with high homology to an artificially generated sequence.

2.1 First Stage: Search for Conserved Sequences (SCS)

Objective. Identify a set of highly conserved sequences W in a multiple sequence alignment S.
Preprocessing. A set of homologue sequences of a protein of interest in a fasta format file (.faa, .fasta), these will be aligned using Clustal Ω a tool for multiple sequence alignment [6, 8, 13, 14]. A fasta file with all sequences aligned will be acquired and afterwards used as input for the first phase of the methodology.
Input data. File in fasta format (.faa, .fasta).
Processing. The file is read, and the alignment's information is stored in $S_{D\bullet l} = \left\{S^i \middle| i = 1, 2, 3, \ldots, D\right\}$, where D represents the number of sequences, l the length of the sequences, $S^i = \left(S^i_j \middle| j = 1, 2, 3, \cdots, l\right)$ i^{th} protein sequence, and j^{th} the alignment's column. The consensus matrix $C(S)$ stores the count for each amino acid by column for each S_j, $C(S) = \left(C(S)^1, C(S)^2, C(S)^3, \ldots, C(S)^L\right)$ where L represents the total length of the alignment, for $C(S)^L = \left\{(S)^L_b \middle| b \in B\right\}$ B

w_i	wa_i	$\alpha SCM(w_i, wa_i)$	$\beta CCM(w_i, wa_i)$	$\gamma HCM(w_i, wa_i)$	$F(w_i, wa_i)$
TGVCV	SGVGV	15.212	30.423	30.423	76.058
VTGVCV	VSGVGV	15.181	30.363	30.363	75.907
TEVVTGVCV	SKVVSGVGV	15.133	30.266	30.266	75.665
VVTGVCV	VVSGVGV	15.160	30.320	30.320	75.799
PNTEVVTGVCV	PNSKVVSGVGV	15.061	30.123	30.123	75.307
EVVTGVCV	KVVSGVGV	15.176	30.351	30.351	75.878
NTEVV	NSKVV	14.878	29.757	29.757	74.392

Fig. 2. Schematic of the txt file containing the condensed results from the ASG stag. Only the best artificial coupling sequence is shown for each highly conserved sequence.

represents the amino acids with its one letter code, $B = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$.

To obtain highly conserved sequences, the length of the substring is defined as Lv which must comply with the following conditions a minimum length of 10 Lvm, a maximum length of 50 LvM, and a conservation value $c = \sum_{j=n}^{m} C(S)^j$ where $n = L^i$ is any position in the alignment $i = 1, 2, \ldots, L - Lvm$ and $m = n + Lv$. The subsequence's length Lv is assigned automatedly using the following approach. Initially the $C(S)$ matrix is iterated from the first column L until $L - Lvm$ with a window equal to Lvm, all resulting subsequences are sorted decreasingly in an array according to its c value.

A subsequence is chosen orderly from this array adding an extra position to the right of the subsequence according to $C(S)$. The new subsequence must comply to the criterion $\Delta c > (5 + (0.02 * Lv))$ where $\Delta c = c_{Lv+x} - c_{Lv}$ and $x = 1, 2, \ldots, 5$. The algorithm continues adding a new position until the criterion is not met or $Lv = LvM$. Finally, a set of several subsequences of size W with varying Lv length and c values will be obtained, order in a decreasing c value fashion where $W = \{w_i | i = 1, 2, 3, \ldots, k\}$ and $w_i = (w_i^p | p_i = \{Lv\})$ p is the number of positions in the sequence w_i with a variable Lv length. Some of the sequences from the set W will be used as input in Subsection 2.2.

Objective Function. $c = \sum_{i=n}^{m} C(S)^i$ where $n = L^i$ is any position in the alignment $i = 1, 2, \ldots, L - Lvm$ and $m = n + Lv$. Obtain a set of sequences with a conservation value c.
Output data. A txt file that stores W in descending order with its c value (Fig. 1.)

2.2 Second Stage: Artificial Sequence Generation (ASG)

Objective. Generate an artificial sequence for each sequence from a subset of W, to obtain a set WA of artificial sequences maximized by its physicochemical compatibility.
Input Data. W, population size, maximum number of generations, crossover probability, mutation probability, and the weight for the parameter in the objective function physicochemical compatibility.
Processing. From W a subset $SW = (w_i | i = 1, 2, \ldots, s)$ of sequences is retrieved, where s is the number of sequences to be retrieved from W and s is defined by the user. The first s sequences from W are retrieved, which will have the best c values.

Algorithm 2 Pseudocode for biological standard mutation

Input: $cross\_population$
Output: $mutated\_population$

$mutated\_population \leftarrow cross\_population$
can_muta generates random number between 1 and P equal to the number of individual to be mutated
these random numbers are saved in ind
for each x contained in ind **do**
    generate a random number between 1 and Lv
    save the random number in $genome$
    Generate a random number between 0 and 1
    Save the random number in $type$
    if $type$ is in the range [0, 0.6) **then**
        mutation by substitution
    if $type$ is in the range [0.6, 0.8) **then**
        mutation by insertion
    if $type$ is in the range [0.8, 1] **then**
        mutation by deletion
    the mutated individual is deposited in $mutated\_population$
return $mutated\_population$

An artificial will be generated for each sequence in SW, this generation of artificial sequences is done by a genetic algorithm (GA) [4]. The GA must be initialized with the following parameters population size P, maximum number of generations G, crossover probability CP, mutation probability MP, and the weight for the parameter in the objective function physicochemical compatibility (PCC).

The GA will repeat the following stages selection, crossover, mutation, passing, and new generation, until the given number of generations previously assigned has been reached. The objective function to maximize the values of these three physicochemical properties with the previously assign weights. The values used to evaluate a given amino acid pair compatibility come from three different matrices: size (SCM), charge (CCM), and hydropathicity (HCM), these are symmetrical matrices that give a compatibility value to each amino acid pair [1].

For each sequence w_i in SW the GA starts with a population of size P of artificial sequences of length $AL = Lv_i$ where i correspond to i in w_i, the sequence composition is generated randomly from B adding a single amino acid until AL length is reached. A set of size P sequences known as WA is acquired, each sequence in WA will be evaluated by the objective function. The selection method can be done by several option such as tournament, roulette wheel, elitism, truncation, or stochastic universal sampling, any one can be used though roulette is chosen by default.

Once the individuals have been selected the crossing process is initiated, also here several option for the crossing can be chosen. The default crossing method is multiple point mimicry, though multiple point, single point, and uniform crossing methods can be selected. The multiple point mimicry method allows a pair of individuals to mimic the composition of the other one, to create two different children with a mixed composition (Algorithm 1).

wa_i	L_i	SP_i
SGVGV	5	17.778
VSGVGV	4	22.222
SKVVSGVGV	1	14.815
VVSGVGV	3	19.048
PNSKVVSGVGV	0	21.212
KVVSGVGV	2	16.667
NSKVV	1	28.889

Fig. 3. Example of the results of the SSA phase that will be stored in a txt.

For the mutation method biological standard mutation is the default option, uniform and standard methods can also be chosen. In the biological standard mutation method, an amino acid mutation propensity matrix is used, this matrix is built considering all single nucleotide insertion, deletion, and substitution that each amino acid coding codon can experiment for each of its three positions. Therefore, each amino acid can only be replaced by an amino acid whose propensity is higher than zero.

Hence, restricting the sequence to a strict evolutionary path that is biologically sound (Algorithm 2). Once this stage is over each individual will be tested with the objective function allowing to make a passing in which all parent sequences will be replace by its offspring, besides the best individual from the parental sequences will be kept if there is no offspring sequence with a better value.

These will breed a new generation of solutions each time the cycle is repeated and on the final generation a population of artificial coupling sequences $WA_f = (wa_i|1, 2, \ldots, P)$ with high compatibility values will be produced for each of the highly conserved sequences evaluated w_i. This set WA_f is saved as txt file and the best wa_i from the set is chosen as result and input for 2.3.

Objective Function. $PCC = max\left(\sum_{j=1}^{AL} F\left(w_i^{Lv_i}, wa_i^{AL}\right)\right)$ where F is an aggregation function of the form $F(a,b) = \alpha SCM(a,b) + \beta CCM(a,b) + \gamma HCM(a,b)$ where $\alpha, \beta, \gamma \in [0,1]$, $\alpha + \beta + \gamma = 1$, SCM, CCM, and HCM are compatibility matrices for each pair of amino acids [1].
Output Data. A txt file containing the following information per column w_i, wa_i, $\alpha SCM(w_i, wa_i)$, $\beta CCM(w_i, wa_i)$, $\gamma HCM(w_i, wa_i)$, and $F(w_i, wa_i)$ (Fig 2).

2.3 Third Phase: Search for Sequences in the Alignment (SSA)

Objective. Locate the column in S where there is a greater similarity with an wa_i.
Input Data. The consensus matrix $C(S)$ that is generated in 2.1 and each wa_i in the txt output file from 2.2.
Processing. The algorithm starts to iterate over $C(S)$ from 1 to $L - AL_i$ in segments of size AL_i for each wa_i looking for the greatest similarity between a segment of $C(S)$ and the wa_i sequence. The position in the alignment L_i is deposited in the variable $max\_homology$, every time a better similarity value is encountered the value of L_i is replaced in the variable $max\_homology$. Once the iteration over all the length L of

Table 1. Input parameters for ASG.

Parameter	**Value**
conservation threshold	80
Population size	100
Maximum number of generations	5000
Crossover probability	80 %
Mutation probability	10 %
α	0.2
β	0.4
γ	0.4

$C(S)$ has finished the following information is saved to a txt file wa_i, L_i, and SP_i that is the value given by the objective function called similarity percentage (SP).

The latter is done for each wa_i, at the end the txt file will contain the best SP_i values and L_i position for each wa_i.

Objective Function. $SP = \max \sum_{j=1}^{AL_i} F2\left(wa_i^j, C(S)^j\right)$ where $F2\left(wa_i^j, C(S)^j\right) = \left\{C(S)_b^j \middle| b = wa_i^j\right\}$

Output Data. A txt file with three columns, in the first one is the artificial sequence, in the second the column's position where the highest similarity percentage is found and in the third the artificial sequence (Fig. 3).

3 Implementation of the Methodology

For the implementation it is first necessary to have a set of proteins of interest, in this case the dataset cliques_066_batch_clique0.faa is used. This dataset is part of a previous work done by our group; it consists of 66 protein homologues clustered by a Bidirectional Best Hits methodology (BBH) [3]. The preprocessing of the data to obtain a multiple sequence alignment is done with Clustal Ω [12].

For the first stage the parameters are the ones described in 2.1, in the second stage it is necessary to set the parameters (Table 1), while the third stage runs as default. Only 9 sequences from the search for conserved sequences algorithm observed the threshold defined. Therefore, the genetic algorithm returned 9 highly compatible sequences, which are searched for similarity segments in the multiple sequence alignment from the dataset cliques_066_batch_clique0.faa (Table 2).

4 Conclusions and Future Work

There is certainty in that the methodology works correctly since the experiments are run several times and the same artificial sequences are found. These artificial sequences had a percentage of compatibility greater than 70%. The weights of α, β, and γ influence greatly the performance of the psychochemical compatibility objective function since each of the parameters from the objective function are mutually exclusive.

Table 2. Results for each stage.

SCS			ASG		SSA	
w_i	Lv	c_i	wa_i	PCC_i	L_i	SP_i
KWPWYVWLLI	10	84.706	DAPAGVAVVI	72.993	1773	31.176
NECVKSQSSRYGFCG	15	83.922	NKGVDSKSSDGGVGG	75.188	1747	36.176
QVDRLITGRLAAL	13	83.823	KVKDVISGDVAAV	75.307	1518	28.959
ECVKSQSSRYGFCGN	15	82.843	KGVDSKSSDGGVGGN	75.188	1748	35.098
YIKWPWYVWLL	11	82.62	GIDAPAGVAVV	72.812	1771	28.877
KVNECVKSQSSR	12	82.23	DVNKGVDSKSSD	75.431	1745	35.049
IEDLLFDKVVT	11	81.684	IKKVVVKDVVS	75.691	1516	26.203
DRLITGRLAALNAFV	15	81.274	KDVISGDVAAVNAVV	75.327	1711	35.196
IKWPWYVWLLI	11	81.15	IDAPAGVAVVI	73.01	1704	29.278

Finding artificial sequence with high compatibility and its subsequent similarity to segments of a protein homologue family can allows us to identify regions of interprotein interaction that might be important for protein function or structure.

The time it takes for the algorithm to run through all stages and return the final output is considerable low, in the order of just a few minutes. Moreover, each stage of the algorithm can be used separately if needed or applied to a different methodology with a separate goal.

4.1 Future Work

Improve the performance of each of the stages to reduce the time taken for the methodology to runs as an all. Explore different options of metaheuristic in the search for highly compatible artificial sequences. Generate a multi-objective paradigm, where each compatibility matrix is separate objective, thus obtaining a set of non-dominated solutions instead of just one with the current aggregation function.

Aknowledgments. The first author is very grateful for the financial support given by the National Council of Science and Technology of México (CONACYT), which allows him to do his postgraduate studies at the Universidad Autónoma de Aguascalientes (UAA) and the authors also acknowledge the support for the PIINF22-3 project, given by UAA, Mexico.

References

1. Biro, J.C.: Amino Acid Size, Charge, Hydropathy Indices and Matrices for Protein Structure Analysis. Theoretical Biology and Medical Modelling, vol. 3, pp. 1–12 (2006). DOI: 10.1186/1742-4682-3-15.
2. Feduchi, E., Blasco, I., Romero, C., Yáñez, E.: Bioquimica conceptos esenciales. Médica Panamericana (2010)
3. Galvis-Motoa, S.I., de Leon, E.P., Marin, E.M., Cuellar-Garrido, D.: Acquisition and Preprocessing of Proteomic Data for Bidirectional Best Hits Methodology: A Study Case in the Coronaviridae Family. Research in Computing Science, vol. 150, no. 9, pp. 1–12 (2021)
4. Haldurai, L., Madhubala, T., Rajalakshmi, R.: A Study on Genetic Algorithm and its Applications. International Journal of Computer Sciences and Engineering, vol. 4, no. 10, pp. 139–143 (2016)

5. Huo, H., Zhao, Z., Stojkovic, V., Liu, L.: Optimizing Genetic Algorithm for Motif Discovery. Mathematical and Computer Modelling, vol. 52, no. 11–12, pp. 2011–2020 (2010). DOI: 10.1016/J.MCM.2010.06.003.
6. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., Mcwilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., Bateman, A.: Clustal W and clustal X version 2.0. Bioinformatics, vol. 23, no. 21, pp. 2947–2948 (2007). DOI: 10.1093/bioinformatics/btm404.
7. Mckee, T., MacKee, J.R.: Bioquímica : Las bases moleculares de la vida. Mc Graw Hill, Access Medicina (2014)
8. Mora-Gutiérrez, R.A., Ramírez-Rodríguez, J., Elizondo-Cortés, M.: Heurística para solucionar el problema de alineamiento múltiple de secuencias. Revista de Matemática: Teoría y Aplicaciones, vol 18, no. 1, pp. 121–136 (2011). DOI: 10.15517/rmta.v18i1.2118.
9. Murray, R.K., Bender, D.A., Botham, K.M., Kennelly, P.J., Rodwell, V.W., Anthony, W.P.: Harper Bioquimica Ilustrada. Mc Graw Hill, (2013)
10. Nelson, D.L., Cox, M.M.: Principles of Biochemistry. Lehninger (2013)
11. Rothlauf, F.: Representations for Evolutionary Algorithms. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion. Association for Computing Machinery, pp. 526–546 (2020). DOI: 10.1145/3377929.3389872.
12. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G.: Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. Molecular Systems Biology, vol. 7, pp. 1–6 (2011). DOI: 10.1038/MSB.2011.75.
13. Zambrano-Vega, C., Cárdenas-Zea, M., Aguirre-Pérez, R.: A Multi-Objective Approach to the Optimization of Multiple Sequence Alignment (MSA). Latin American Journal of Computing, vol. 3, no. 1, pp. 43–51 (2016). DOI: 10.5281/zenodo.5748527.
14. Zambrano-Vega, C., Nebro-Urbaneja, A., Aldana-Montes, J.F.: Metaheurísticas de optimización multiobjetivo aplicadas a la inferencia filogenética y al alineamiento múltiple de secuencias. Universidad de Málaga (2017)