Journal of Phylogenetics & Evolutionary Biology

Giuliani et al., J Phylogen Evolution Biol 2013, 1:2
DOI: 10.4172/2329-9002.1000111

Research Article Open Access

Amino-Acid Correlated Mutations inside a Single Protein System: A New Method for the Identification of Main Coherent Directions of Evolutive Changes

Alessandro Giuliani[1], Roberto Bruni[2], Massimo Ciccozzi[2*], Alessandra Lo Presti[2], Michele Equestre[3], Cinzia Marcantonio[2] and Anna Rita Ciccaglione[2]

[1]*Department of Environment and Health, Istituto Superiore di Sanità, Rome, Italy*
[2]*Department of Infectious, Parasitic and Immunomediated Diseases, Istituto Superiore di Sanità, Rome, Italy*
[3]*Department of Cell Biology and Neurosciences, Istituto Superiore di Sanità, Rome, Italy*

Abstract

The need of giving rise to a stable and soluble protein system generates constraints that limit the mutation space by imposing a co-variation structure across different residues. While protein scientists widely use this property in order to predict protein-protein interaction and peptide-receptor pairing, there is no equivalent interest to make use of mutation correlation structures to get information inside single protein systems. Here we present a methodological essay that, using a statistical approach typical of medicinal chemistry, faces the problem to locate 'mutational correlation units' in a viral RNA polymerase. These 'units' are invisible to ordinary sequence alignment methods and can be important in virus characterization and vaccine developments as well as for evolution theorists.

Keywords: Co-evolution; Structural contacts; Mutational units; HCV genotype; HCV strains; Epidemiology

Introduction

We were happy to read the two related papers, recently appeared on a recent issue of Nature [1,2] putting evolution in the perspective of protein science. While protein scientists routinely used since years the co-evolution degree of protein systems so to infer protein interacting pairs [3], evolutionary biologists were still linked to an old idea of 'epistasis' (between genes functional correlation producing a correlation in their mutation dynamics) as a relatively marginal effect. This gap prevented the adaptation of evolution theory to the recent developments of biology, with the embarrassing presence of 'an-elephant-in-the-room': the network character of biological regulation at all the scales from single molecule to organs. Tompa and Rose [4] clearly stated that 'The central biological question of the 21st century is: how does a viable cell emerge from the bewildering complexity of its molecular components?' pointing to the accurate wiring of protein-protein interaction network as the main issue of nowadays theoretical biology. Now evolution theory joins the game as clearly indicated by Wagner [1] statement "Amino-acid substitutions will persist, on an evolutionarily relevant timescale, only when the 'correct' amino-acids are present elsewhere in the protein".

This statement, beside the theoretical relevance for evolution theory, has a very important applicative consequences, especially in the case of virus evolution that, thanks to its accelerated pace, could generate new variants with different pathological implications.

This was the case of Hepatitis C Virus (HCV) were was recently recognized, inside the so called genotype 1b strain, the presence of two variants characterized by two different versions of the same protein (NS5B, the viral RNA polymerase), one characterized by a cystein in the 316 position (C-316) and one by an asparagine (N-316).

The interest in this polymorphism stems from the different susceptibility of the two variants to in vitro therapeutic treatment with a drug called HCV-796 (benzofuran carboxamide) which is a potent Non-Nucleoside Inhibitor (NNI) of the enzyme which binds near the catalytic site (GDD motif; guanine 317, aspartic acids 318 and 319). The *in vitro* analysis of HCV replicons containing the C316N mutation showed that this variant displayed a 26-30-fold-reduced susceptibility to HCV-796. Nevertheless, the substitution was never selected after treatment with HCV-796 *in vitro* or in clinical trials [5]. This observation is apparently in contrast with the wide diffusion of C316N variant among the natural isolates (114 out 309 HCV 1b isolates found in GeneBank) which suggests a high level of fitness of the N316 virus. One possible explanation is that the emergence of C316N mutation during treatment may be hampered by the need of a concurrent selection of compensative mutation in one or more positions which would allow a sufficient level of efficiency of NS5B polymerase. These amino acid variations might be already present in N316 natural isolates.

Moreover, the markedly different geographical distribution of C and N isolates (C isolates are much more frequent in Western countries, while N isolates mainly come from Eurasia) makes the analysis of this protein system a perfect benchmark for the presence of a 'mutational pattern' relative to amino-acid residues different from the target mutation at 316 location. The discovery of such co-varying locations could both open new avenues for alternative therapeutic strategies and shedding light on the dynamics of virus evolution. We tried and solve the problem of co-evolution clusters inside the same protein with a combined supervised / unsupervised statistical approach inspired to Quantitative Structure Activity Relationships (QSAR) methods, widely used in both medicinal chemistry [6] and toxicology [7] in which the discrimination of a set of molecules in two or more activity classes is faced by the analysis of chemico-physical properties of the compounds.

We were able to demonstrate how this approach is able to attain an almost perfect discrimination of the two sets and to single out 'mutation clusters' of residues inside the studied protein.

***Corresponding author:** Dr. Massimo Ciccozzi, Department of Infectious, Parasitic and Immunomediated Diseases, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161, Rome, Italy, Tel: +39-06-49903187; E-mail: massimo.ciccozzi@iss.it

Received June 22, 2013; **Accepted** July 19, 2013; **Published** July 22, 2013

Citation: Giuliani A, Bruni R, Ciccozzi M, Lo Presti A, Equestre M, et al. (2013) Amino-Acid Correlated Mutations inside a Single Protein System: A New Method for the Identification of Main Coherent Directions of Evolutive Changes. J Phylogen Evolution Biol 1: 111. doi:10.4172/2329-9002.1000111

Material and Methods

Our raw material came from viral isolates relative to 309 different patients, all belonging to the same 1b genotype. Each one of these isolates showed the presence of only one by far major represented viral variant, endowed with a specific version of NS5B, this allowed us to group the initial data set into two classes correspondent to the C-316 (195 units) and N-316 (114 units) isolates. Each isolate is coded as a 590 components row vector whose elements correspond to the subsequent aminoacid residues along the chain with the exception of position 316 carrying the target mutation. Only 291 positions out of 590 demonstrated a different from zero variability (non-strictly-conserved across the considered population) . The task was to correctly classify each single isolate (sequence) into the C-316/N-316 categories on the basis of the information coming from the 291 positions along the sequence that potentially carry information about the nature of aminoacid residue in position 316. This task was accomplished by a linear discriminant analysis as applied to the amino-acid location coded in terms of Miyazawa and Jernigan et al. [8] hydrophobicity.

Having solved this problem with a good precision, we tried and detect in an unsupervised way the presence of 'mutation clusters' in the protein sequence pointing to groups of amino-acids going along a common mutation path: the ability to recover (without explicit indication) the C-316/N-316 by an unsupervised approach (Principal Component Analysis (PCA)), gave a proof-of-concept of the general relevance of the method to locate common evolution patterns probably linked to hidden structural and functional units in the protein system.

Results

Before entering the discriminant analysis step, the pairwise Pearson correlation coefficients between hydrophobicity value at position 316 and hydrophobicity values relative to the 291 other locations showing some variability were computed. Given position 316 has only two possible hydrophobicity values correspondent to the two C and N residues with hydrophobicity values of 7.93 and 3.71, respectively. A positive correlation with the generic position (i) implies C-316 mutants tend to display a more hydrophobic residue in position (i) with respect to N-316, while a negative correlation points to the opposed pattern and a zero value correlation indicates the lack of any relevant co-variation between the (i) and 316 positions. Thus, the entity of correlation with position 316 (a316) measured by Pearson correlation coefficient corresponds to the strength of co-variation of each residue with the target mutation in the analyzed set. This correlation stems from the need of generating a stable and viable structure as a whole, thus the specific mutation in 316 positions must 'accommodate' itself with concerted changes at different locations.

Table 1 reports the basic statistics for the 291 Pearson correlation values of the different sequence locations with position 316 as computed along the 309 statistical units (different isolates).

Mean	Standard Deviation	Q90	Q10	Min	MAX
-0.00083	0.126113	0.090442	-0.105563	-0.520539	0.750214

Table 1: Reports the basic statistics for the 291 Pearson correlation values of the different sequence locations with position 316 as computed along the 309 statistical units (different isolates). Q90 and Q10 correspond to the 90% and 10% Quartile of the distribution.

Observed / Predicted	Pr(C-316)	Pr(N-316)
Ob(C-316)	**190**	5
Ob(N-316)	3	**111**

Table 2: The correct predictions were the bolded values in which predicted class Pr(#) coincides with observed class Ob(#).

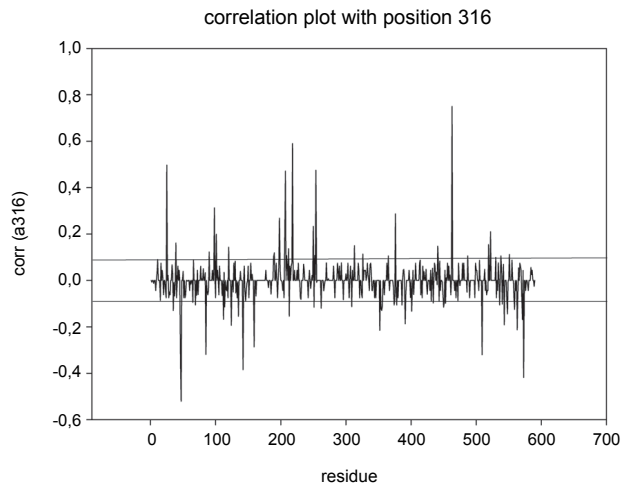

Figure 1: The figure depicts the correlation coefficients with position 316 (corr (a316)) as histograms along sequence positions (on the X axis), while the two parallel lines correspond to the Q90 and Q10 values. It is worth noting the emerging of few very high (or equivalently very low) peaks corresponding to locations whose mutation pattern is correlated (in the MJ hydrophobicity space) with target mutation.

As evident in Table 1, the average correlation with position 316 is extremely low (practically 0), but the high variability of the coefficients distribution as measured by Standard Deviation, Q90 (the value greater than 90% of the observation) vs. Q10 (value greater than 10% of observations) and Range (minimum (min) vs. maximum (MAX) difference) tells us that there are some very relevant 'peaks' constituted by few specific residues showing a relevant co-variation with position 316. This is evident in Figure 1 reporting the correlation coefficients with position 316 (corr (a316)) as histograms along sequence position (on the X axis), while the two parallel line correspond to the Q90 and Q10 values.

In order to go from a purely descriptive to a predictive model, we based a stepwise discriminant analysis upon the 75 residues mostly correlated with position 316 so to generate a predictive model of the pertaining of each isolate to the C-316 or N-316 class. The stepwise discriminant analysis tries to create the most parsimonious, albeit endowed with the maximal discriminative power, model by progressively introducing explanatory variables (hydrophobicity values relative to different locations out of the 75 previously selected) until a maximum discrimination between the two classes is achieved. The procedure selected 11 locations to build the model namely (in order of discrimination power): 464, 218, 159, 510, 313, 47, 254, 98, 574, 207, 377 (Figure 2, where the selected locations are marked by an asterisk).

The model allowed for an almost perfect identification of the phenotypes with around 97% accuracy, Table 2 reports the so called 'confusion matrix', i.e. the two entries contingency table reporting in rows and columns the observed and predicted categories, respectively.

Now we are in the position of safely saying that a single mutation in a specific location along a protein sequence is strictly correlated with an array of other mutations in different parts of the molecule (not necessarily directly interacting with the targeted mutation), so verifying amino-acid co-evolution INSIDE a given molecule. The emerging of a

single main isolate from a huge virus population infecting a patient, constituting a sort of 'accelerated evolution system' will give a strong support to the novel consideration of epistasis in evolutionary studies [1,2] at the microscale of a single protein systems.

In order to check if locations along the sequence whose respective mutational spaces are each other correlated can be detected even with an unsupervised approach, i.e. without explicitly inserting into the goal function (like is the case with discriminant analysis) the maximal separation between *a priori* selected groups, we applied on our data matrix a PCA. The aim was to check if an unsupervised method was able to spontaneously give rise as first principal component (corresponding to the main order parameter shaping the among sequences variation in terms of correlated mutational load) the 'hidden' mutation cluster made of the locations previously selected by supervised strategy.

In other words we want to check the congruency between the first principal component of the whole set and the results of discriminant analysis.

It is important to stress that PCA is totally blind to the separation of the data set into two C-316 and N-316 classes and its only goal is to project the multidimensional raw data set spanned by different hydrophobicity values into a lower dimensional space spanned by new composite variables called principal components [9]. The principal components are each other orthogonal by construction, thus each component corresponds to an independent mutational path and its correlation coefficients (loadings) with the original variables (locations) measure the entity of the involvement of each residue on the specific mutation path. The component space is maximally parsimonious, so the components are extracted in order of 'explained variation', thus the first component will correspond to the most relevant flux of variation, the second to the second most relevant mutational flux orthogonal to the first one and so forth. The procedure can be stopped when reaching the so called 'noise floor' corresponding to a plateau of cumulative explained variation that, in the present case, can be intended as the free, un-constrained random mutation flux of each residue [9,10].

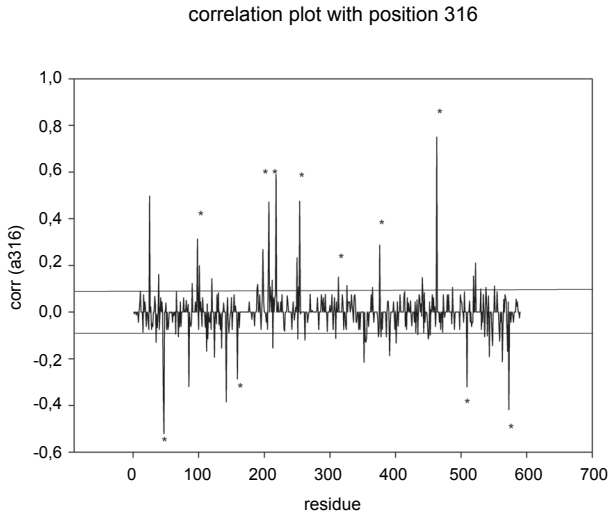

Figure 2: The figure reports, marked by an asterisk, the locations that were selected by Stepwise Discriminant Analysis for discriminating C-316 and N-316 groups. These locations roughly correspond to the peaks of Fig.2 but they are not identical, given the procedure does not select mutually correlated regressors (locations) because they do not add new information useful for discriminant function. The procedure selected 11 locations to build the model namely (in order of discrimination power): 464, 218, 159, 510, 313, 47, 254, 98, 574, 207, 377.

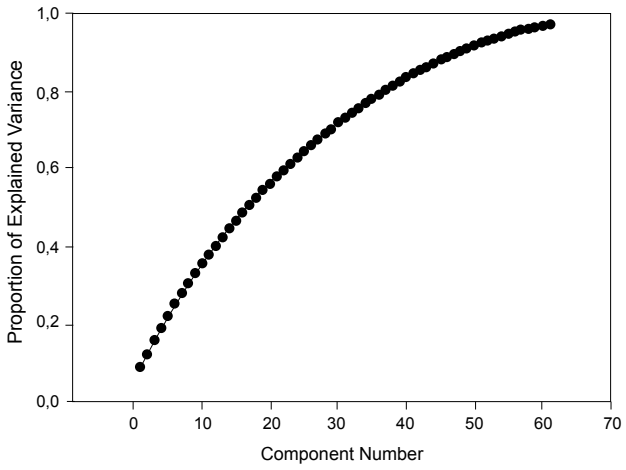

Figure 3: The cumulative proportion of explained variance at increasing number of components is reported. The graph tends to a plateau after which no amelioration in explanation power is possible by the increasing complexity of the model. Selecting a number of components identical to the number of variables generates a trivial complete explanation correspondent to a geometrical rotation of the initial data set.

In Figure 3 is reported the distribution of proportion of variance explained along the components.

As it is evident from Figure 3, the mutational space of NS5B shows a strong internal correlation structure with approximately 30-40 independent 'mutation paths' out of the 291 residues showing some polymorphism (and 591 total residue numbers). The estimate of the number of mutation paths is based on the fact the total proportion of variation present in the data set explained by principal component solution start to reach a plateau around the PCA solution at 30-40 components, i.e. using 30-40 independent 'blocks of covariant residues'. It is important to keep in mind each residue can pertain to more than one 'mutation path' and that we do not expect the 'definitive' solution can ever reach the total variance explained condition, given in principle we can imagine that there is room for 'singular' random drift at each location. Nevertheless the above analysis can give us some indication about the existence of possible 'modules' inside the protein that are sensitive to different selection biases.

We can prove the above interpretation thanks to the fact in this particular case we purposedly 'inserted' such selection bias in the form of C-316/N-316 categories, thus, given position 316 was not inserted in PCA computation, if our model is correct, we expect the most relevant (mostly loaded, both with positive and negative correlation) variables on the most relevant 'mutation path' (PC1) present in our data set should be largely coincident with the 11 most discriminative positions highlighted by discriminant analysis. This was actually the case as evident in Table 3 reporting the 20 variables mostly correlated with PC1, bolded values correspond to the locations shared with supervised analysis 'most discriminant' positions.

Table 3 shows a good concordance between residues mostly loaded on PC1 and the residues selected as the 'most informative' by discriminant analysis, the most cogent proof of the consistency between the most relevant mutation path suggested by data and the *a priori* inserted selection bias is the strong Pearson correlation coefficient between the probability to be a member of C-316 class as estimated

Residue	Corr (PC1)
A25	0.782
A47	**-0.746**
A207	**0.709**
A464	**0.677**
A198	0.642
A218	**0.633**
A85	-0.586
A142	-0.553
A250	0.551
A101	0.499
A254	**0.469**
A251	-0.460
A98	**0.392**
A392	-0.352
A124	-0.350
A135	-0.300
A114	-0.297
A574	**-0.284**
A39	0.255
A213	-0.238

Table 3: Reports the 20 variables mostly correlated with PC1, bolded values correspond to the locations shared with supervised analysis 'most discriminant' positions.

by the discriminant function (P(C-316)), and PC1 scores computed over the 309 isolates and that was equal to r=0.8438 (p<0.0001). This evidence was supplemented by the highly significant statistical differences between PC1 scores relative to C-316 and N-316 class (t-value=19.9, p<0.0001) corresponding to an average PC1 value of 0.62 for C-316 (Std.Dev.=0.27) and of -1.06 for N-316 (Std.Dev.=0.88).

Having demonstrated the possibility to predict in an unsupervised way the most relevant mutation fluxes present in a protein data set is important to check if the same goal could be achieved by a more classical bioinformatics procedure, i.e. the construction of a hierarchical classification tree of our sequences on the basis of their homologies.

In Figure 4 we report two classification trees: in panel 4a the classification obtained taking explicitly into consideration position 316, while in panel 4b is reported the sequence homology classification of the same isolates without making use of target mutation information. Panel a) is thus coincident with the 'data set as it is' and simply mirrors the purposedly inserted selection bias in which an almost equally distributed C-316/N-316 data set was built, in other words Panel 4a is nothing more nothing less than the image in light of our inclusion criteria. On the other hand panel b) is the one that must be compared with the results of our proposed method, being based on the same kind of primary information, i.e. the primary structures of the analyzed proteins with the exception of target mutation.

If we consider the disposition of the entire set of 309 isolates we can easily appreciate the mutation space analyzed is very narrow: the reported bar corresponding to an average of 0.04 substitutions per site and, when compared with the actual length of the branches of the trees. it allows us to appreciate we are dealing with very similar sequences. This is consistent with the fact we are analyzing isolates of the same genotype.

The differential coloring of the two subgroups tells us (Panel a) that the C-316/N-316 systematic bias is faithfully registered by the localization of red and blue leafs in two separate domains of the trees. When shifting to the purely contextual information without taking into account the target mutation (Panel b), this discrimination is much less evident with N-316 group scattered around the tree.

On the contrary, if we consider the probability assigned by discriminant analysis of being part of group N (P(N-316)) we observe a clear-cut bimodal distribution with two modal classes correspondent to a first class with P(N-316) near zero made only of C-316 isolates and a second class made only of N-316 sequences with a P(N-316) near to unity, only very few sequences have an uncertain status, some of them corresponding to the wrong classifications reported in Table 2 (Figure 5).

The above results point to the fact the proposed method is able to discovery very fine details of protein modification that cannot be highlighted by traditional sequence homology approaches.

Discussion

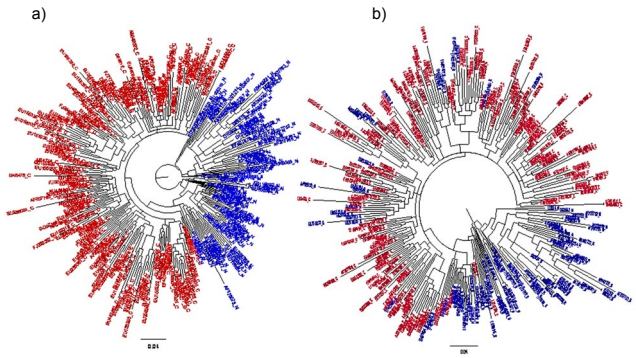

Figure 4: Maximum likelihood phylogenetic trees relative to the 309 HCV NS5B genotype 1b sequences. C-316 sequences are colored in red, while N-316 sequences are colored in blue. The scale bar at the bottom of the tree indicates 0.04 amino-acid substitution per site. Panel a) reports the full-information tree (target mutation inserted in the tree generation), while Panel b) was computed without taking into account position 316.

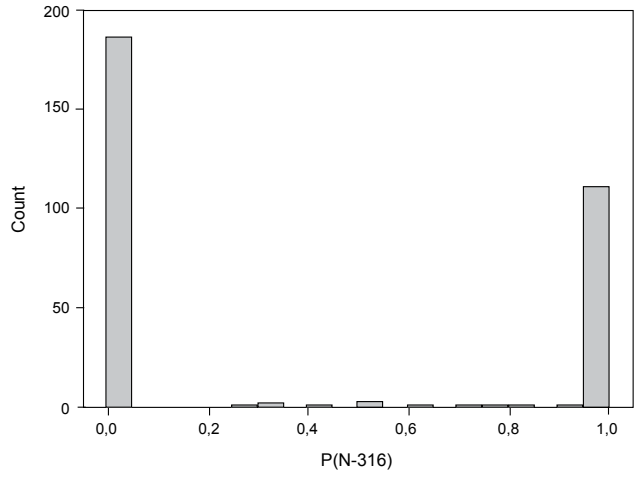

Figure 5: Distribution of P(N-316) variable corresponding to the probability to pertain to the N-316 group as estimated by Discriminant Analysis. It is worth noting the neat separation of the proteins into two non-overlapping clusters of near to zero and near to unity probability values.

The discovery of common fluxes of mutation by means of regressive approach inspired by QSAR analysis is made simple by the natural order of aminoacid disposition along the sequence: the classical scheme of a set of chemico-physical descriptors defining a given drug candidate in medicinal chemistry becomes, in the case of protein sequences, the array of one (or more) chemico-physical properties of the constituent aminoacids. The use of equally length sequences makes the procedure easier; nevertheless the transformation of the raw data set by the action of autocorrelation of the studied property at different distances on the sequences [11] allows extending the procedure to different length sequences.

The convergence of supervised (discriminant analysis) and unsupervised (PCA) approaches toward a common result is a strong proof-of-concept of the relevance of the proposed method for exploratory purposes.

Like any statistical investigation, the relevance of the obtained results is critically dependent from the data set features, in this case our work was made easier by the presence of a strong a priori selection bias correspondent to the C-316/N-316 categorization that worked as order parameter, in the case of an hypothesis generating (and not an hypothesis testing setting like ours) exploratory study the presence of a large data set of randomly selected sequences could be a good starting point to individuate what we called 'correlated mutation' groups of aminoacids.

The discovery of such clusters can be useful for practical application, as matter of fact in our case this discovery tells us that trying to base the therapeutic experimentation (or analogously the search for a vaccine) making use of a variant in which the only difference with 'wild-type' is at position 316 is not a good strategy because it does not mirror the natural co-occurrence of target mutation with other correlated changes in the protein.

On a more theoretical ground, the fact the C-316/N-316 polymorphism has a geographical basis gives an evolutionary meaning to our finding giving a demonstration of the presence of a strong canalization of allowed evolution changes even at the single protein microscale and not only at the level of macroevolution . The need to taking into account not only the qualitative information about the different residues at different locations but their chemico-physical protein is a strong indication toward the need of a strong integration of different fields of science as evoked by Wagner [1] connecting genetics to the 'inner life of proteins'. The next step should be the study of the relation of the 'mutation clusters' with the structural location of the residues in the protein 3D structure, we are actually following this path that in any case deserves a much wider experimentation involving many different protein systems.

References

1. Wagner GP (2012) Genetics: The inner life of proteins. Nature 490: 493-494.
2. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. Nature 490: 535-538.
3. Goh CS, Cohen FE (2002) Co-evolutionary Analysis reveals insights into protein-protein interaction. J Mol Biol 324: 177-192.
4. Tompa P, Rose GD (2011) The Levinthal paradox on interactome. Protein Sci 20: 2074-2079.
5. Howe AY, Cheng H, Johann S, Mullen S, Chunduru SK, et al. (2008) Molecular mechanism of hepatitis C virus replicon variants with reduced susceptibility to a benzofuran inhibitor, HCV-796. Antimicrob Agents Chemothe 52: 3327-3338.
6. Hansch C, Leo A (1995) Exploring QSAR. Washington: American Chemical Society.
7. Hansch C, Kim D, Leo AJ, Novellino E, Silipo C, et al. (1989) Toward A Quantitative Comparative Toxicology of Organic Compounds. Crit Rev Toxicol 19: 185-226.
8. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18: 534-552.
9. Benigni R, Giuliani A (1994) Quantitative modeling and biology: The multivariate approach. Am J Physiol 266: R1697-R1704.
10. Preisendorfer RW, Mobley CD (1988) Principal Component Analysis in Metereology and Oceanography. Develop Atmosph Sci 17 Amsterdam: Elesevier.
11. Tong J, Che T, Liu S, Li Y, Wang P, et al. (2011) SVEEVA Descriptor Application to Peptide QSAR. Arch Pharm (Weinheim) 344: 719-725.