W412–W419 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 7 June 2022*
https://doi.org/10.1093/nar/gkac481

iBIS2Analyzer: a web server for a phylogeny-driven coevolution analysis of protein families

Francesco Oteri, Edoardo Sarti, Francesca Nadalin* and Alessandra Carbone*

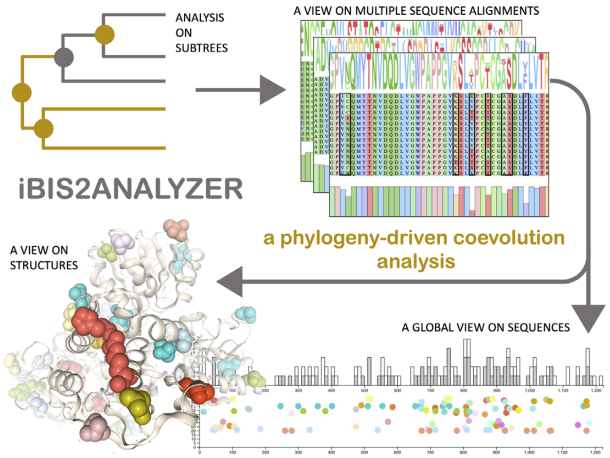
Sorbonne Université, CNRS, IBPS, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France

Received March 23, 2022; Revised May 20, 2022; Editorial Decision May 23, 2022; Accepted May 25, 2022

ABSTRACT

Residue coevolution within and between proteins is used as a marker of physical interaction and/or residue functional cooperation. Pairs or groups of coevolving residues are extracted from multiple sequence alignments based on a variety of computational approaches. However, coevolution signals emerging in subsets of sequences might be lost if the full alignment is considered. iBIS2Analyzer is a web server dedicated to a phylogeny-driven coevolution analysis of protein families with different evolutionary pressure. It is based on the iterative version, iBIS^2, of the coevolution analysis method BIS, Blocks in Sequences. iBIS^2 is designed to iteratively select and analyse subtrees in phylogenetic trees, possibly large and comprising thousands of sequences. With iBIS2Analyzer, openly accessible at http://ibis2analyzer.lcqb.upmc.fr/, the user visualizes, compares and inspects clusters of coevolving residues by mapping them onto sequences, alignments or structures of choice, greatly simplifying downstream analysis steps. A rich and interactive graphic interface facilitates the biological interpretation of the results.

GRAPHICAL ABSTRACT



INTRODUCTION

Coevolving residues in a protein structure, possibly a complex, correspond to groups of residues whose mutations have arisen simultaneously during the evolution of different species, and this is due to several possible reasons involving the three-dimensional shape of the protein: functional interactions, conformational changes and folding.

In the last twenty years, a particular focus has been drawn to the development of methods identifying pairs of coevolved residues in contact within a protein with the highest accuracy (1–9). The recent impressive progress made by AlphaFold2 on the prediction of the three-dimensional structure of proteins with atomic accuracy relies on coevolution signals (10,11). Besides physical contact (12,13), coevolved residues have been demonstrated to play a crucial role in allosteric mechanisms (1,3,14), to maintain short paths in network communication and to mediate signaling (15). Moreover, coevolution analysis of highly conserved proteins, such as the Amyloid beta peptide playing an important role in Alzheimer's disease, and families of very few sequences such as the ATPases characterized by conserved motifs in divergent sequences, allowed to highlight that co-

*To whom correspondence should be addressed. Tel: +33 1 44 27 73 45; Fax: +33 1 44 27 73 36; Email: alessandra.carbone@lip6.fr
Correspondence may also be addressed to Francesca Nadalin. Email: francesca@ebi.ac.uk

evolving protein fragments, and not only residues, are indicators of important information explaining: folding intermediates, peptide assembly, key mutations with roles in genetic diseases, distinguished subfamily-dependent motifs (16–18), conformational changes (19), and secondary mutations in viruses induced by viral adaptation to drugs (20).

iBIS2Analyzer is a web server dedicated to a phylogeny-driven coevolution analysis of protein families characterized by different evolutionary pressure. It can analyze a few dozens up to some thousands of sequences and identifies coevolution signals which might be specific to protein subfamilies. Starting from the multiple sequence alignment (MSA) and its associated phylogenetic tree T, iBIS2Analyzer employs a bottom-up iterative approach to select subtrees in T and find coevolution signals with BIS^2, a fast version of BIS (16,17), specifically designed to handle few and/or conserved sequences. The key idea is that highly related sequences may contain evolutionary information that is lost in the global MSA, where local signals might be highly diluted. The iterative nature of iBIS2Analyzer makes it highly adjustable to protein families with different amino acid variability and size, filling the gap between existing statistical and combinatorial methods of coevolution analysis. Its flexible design facilitates the analysis of coevolutionary signals in sequences, alignments and structures. By applying iBIS2Analyzer to concatenations of MSAs associated to multiple proteins, the user can explore clusters of positions both within and between proteins.

METHODS

The BIS^2 and iBIS^2 algorithms

The web server works in two different modes by calling two distinguished but related algorithms, 'BIS^2' and 'iBIS^2'. The first mode runs BIS^2, a combinatorial method specifically designed to find coevolution signals in small or conserved MSAs (16,17). The web server BIS2Analyzer provided the first online access to the BIS^2 algorithm, described in detail in (18).

The second mode runs iBIS^2, an iterative procedure that explores the topology of the phylogenetic tree associated to the input MSA and runs BIS^2 independently on selected subtrees; it is designed to analyze larger MSAs, possibly comprising thousands of sequences. The goal of iBIS^2 is to find coevolution signals within subfamilies, identified by subtrees, which are diluted in the full tree. More precisely, iBIS^2 employs a bottom-up iterative approach to select and analyse subtrees of the phylogenetic tree. Coevolution clusters are computed on different subtrees independently to explore (i) coevolution signals of subfamilies, identifiable in subtrees, which are diluted in the full tree, (ii) coevolution signals occurring in non-overlapping subtrees, (iii) coevolution signals that are conserved in a cascade of subtrees, one contained within the other, and that are extended with coevolving residues found in the smaller trees. iBIS2Analyzer is designed to facilitate the exploration of these multiple coevolution signals.

Strategy for subtree selection. Let T be the input tree. iBIS^2 determines first a list of eligible subtrees of T that contain a number of leaves in a range $r = [n, N]$ ($n = 20$ and $N = 300$, by default) and differ from one another by at least K leaves ($K = 5$, by default). To do this, it uses an iterative procedure. It considers the list $\mathcal{T}$ of all subtrees of T whose size is within the interval r. At iteration $i = 1$, it selects a maximal collection $\mathcal{S}_1$ of pairwise disjoint subtrees in $\mathcal{T}$ of minimal size. At iteration $i + 1$, it selects the largest collection $\mathcal{S}_{i+1}$ of subtrees of T such that 1. given any two selected subtrees T', $T'' \in \mathcal{S}_{i+1}$, their intersection is empty, and 2. the maximal intersection of a subtree $T' \in \mathcal{S}_{i+1}$ with some $T'' \in \mathcal{S}_j$, where $j \leq i$, contains at least K leaves. This means, for instance, that if a subtree T' has two immediate subtrees T'_1 and T'_2 of 2 and 40 leaves respectively, and that T'_2 is in $\mathcal{S}_i$, a $K = 5$ will not allow to select T' since it contains only 2 more new sequences than T'_2. The iteration ends when no more subtrees can be selected. Note that large subtrees with a leaf count of $>N$ cannot be selected.

A second round of filtering is applied to the list of selected subtrees and keeps only those that contain at least 2 non-conserved positions (up to exceptions, see (16)) in the associated MSA. By default, sequences in a subtree are aligned as in the input MSA; with the option 'Realign', they are realigned for each subtree (see 'The Submission Page' below).

Starting from the smallest subtrees in the final list of selected subtrees $\mathcal{S}$, iBIS^2 iteratively considers the subtrees in $\mathcal{S}$ and runs BIS^2 on each of them following their order. iBIS2Analyzer reports all coevolution signals obtained in the analysis of all the subtrees. The range $r = [n, N]$ and the value K are set by default by iBIS2Analyzer but can be changed by the user. Given an alignment, only non-conserved positions are selected for coevolution analysis.

The Submission Page

The web server provides a job *Submission Page*. Its basic usage is intuitive. To have a glimpse on the type of input required, sample inputs can be loaded for intra- and inter-protein coevolution analysis. For information on how to customise the default behaviour, a detailed documentation is accessible online at the *Documentation Page*.

Details on the input data. iBIS2Analyzer accepts as input a MSA in FASTA format, either copy-pasted or uploaded as a file. There is no restriction for sequence names but the display might be shortened and the user can visualise the entire name by hovering over the short version with the cursor. Before the job is submitted, the MSA undergoes a syntactic check controlling the expected format. Error descriptions help the user fixing the syntax. iBIS2Analyzer also accepts previous job files as input, allowing the user to view past runs.

iBIS2Analyzer default parameters and options. iBIS2Analyzer suggests to use iBIS^2 by default. For small or very large sets of sequences, the user can adapt the minimum and maximum numbers n, N of sequences in a subtree ($n = 20$ and $N = 300$, by default) and the minimum number K of new sequences contained in a selected subtree ($K = 5$, by default). For instance, for small sets of sequences, one may want to analyse all subtrees of the input tree containing at least 5 and at most 200 sequences, setting $K = 1$, $n = 5$ and $N = 200$.

BIS^2 can be run by checking the corresponding option. Parameters specific to iBIS^2 or BIS^2 analyses are described in the Documentation Page. Details for parameters specific to BIS^2 are found in (16,18).

The user can either provide a phylogenetic tree in NEWICK format (either copy-pasted or uploaded as a file), or run PhyML (21), FastTree (22) or BioNJ (23). FastTree is applied in its double-precision version FastTreeDbl; it takes a MSA as input and returns an unrooted tree. BIONJ is run by first computing the distance matrix, based on the Jones-Taylor-Thornton distance model (24) with Protdist, from PHYLIP version 3.696 (25). SeaView is used to re-root the tree (26) for both FastTree and BioNJ (with `seaview -reroot -root_at_center`).

The option 'Realign', by default, is not activated. This option realigns the sequences in each subtree considered for analysis. Since alignments can change considerably within each subtree, no mapping of clusters identified in a subtree is pushed back to the consensus sequence of the global MSA. However, the clusters computed in a given subtree can be analysed in the corresponding tab of the sub-MSA and mapped in the 3D structures.

The Progress Page

After submitting the job, the user can follow the progress of the job through a description organized in five main steps (for iBIS^2: Job setup, Tree construction, Subtree selection, iBIS^2 execution, Finished; for BIS^2: Job setup, Tree construction, BIS^2 execution, CLAG execution, Finished). The icon for each step will be colored green, yellow and red depending on the progress status: completed, active, to do, respectively. Below the progress bar, the distance tree constructed for the input sequences displays the progress of iBIS^2 step #4 by coloring completed nodes green, nodes under consideration yellow and nodes to be considered red.

The Results Page

The *Results Page* shows the collection of coevolution clusters found by the algorithm of choice (iBIS^2 or BIS^2). iBIS2Analyzer names each cluster by a number used to refer to the cluster across the three main visualization panels (Figure 1): the Clusters Panel, the MSA Panel, and the Tree Panel. It also gives access to the Clusters Tab window providing detailed descriptions of all coevolution clusters. The cross-talk among panels is mediated by a number of control buttons whose effect is propagated automatically. They allow the user to create a working environment through the specification of groups of clusters to analyse (by hand or based on thresholds), the restoration of the default session for the analysis of all clusters, the visualization of the clusters on the MSA, the visualization of subtrees and their corresponding sub-MSAs and clusters, the visualization of the clusters on 3D structures. If iBIS^2 is used, the user can display different Clusters Panels for different subtrees. See Figure 1 for an example of Clusters Panel analysis dedicated to a specific subtree. The Clusters Panel associated to the entire tree, called '1', is always displayed.

The Clusters Panel. A Clusters Panel shows the positions of each cluster along the (sub-)MSA consensus sequence through two main visual representations:

i. the *Zoom Plot* is an interactive histogram displaying the count of all coevolution signals found along the MSA. By displaying all clusters in once, the user has a global vision of the density of coevolving positions throughout the MSA length, and possibly, of the regions of the protein that are affected or not by coevolution signals. The bars of the histogram are filled dynamically according to the current selection (see button 'Select on position' in Figure 1). By default, all coevolving positions are visible, and the histogram bars are filled to 100%. The user can zoom into a portion of the histogram to select a region of the MSA. The filled portion of the bars is proportional to the count of coevolution signals from clusters having at least a position within the selected region. The Clusters Plot (see below) will automatically adapt to the selection. Positions in selected clusters are highlighted in the MSA Panel with black borders.
ii. the *Clusters Plot* displays all positions in the identified clusters as colored dots. Positions in the same cluster have the same color and each cluster has a different color. Clusters are stacked according to their numerical name (y-axis). For instance, in Figure 1, there are 65 clusters in all as shown on the y-axis and two of them, circled in black, are identified in subtree '98'. By hovering the mouse over the colored dots, the user can access the position in the (sub-)MSA consensus sequence, the cluster name, and the subtree ID associated with the selected cluster when iBIS^2 is used.

The Clusters Tab. In this pop-up window, accessible on the right of the Results Page (Figure 1), clusters can be ordered by different parameters (*P*-value, dimension, symmetric/Ssymm and environmental/Senv scores; if iBIS^2 is run, subtree name and iteration level are also given) and be selected for visualization. Also, the user can manually assign the colour to a cluster by clicking on its coloured square, and can switch it on/off. Changes are automatically propagated over the whole interface. The Clusters Tab gives also access to the amino acid pattern, through the button 'Show Patterns' (see Figure 1).

The *P*-value score is computed with a Fisher test on a diagonal matrix, where the elements of the diagonal represent the coevolution pattern satisfied by all positions in a cluster (18); see the Documentation Page for an example. Symmetric and environmental scores vary in the interval [0,1] and are computed by the clustering algorithm CLAG (27). They express the degree of 'similarity' of coevolution of positions in a cluster with respect to all other positions in the alignment. In particular, scores equal to 1 correspond to a cluster where all positions show an identical coevolution pattern with all other positions in the alignment. High scores guarantee the confidence in a cluster; note that iBIS2Analyzer outputs only clusters with both scores >0.5.

The MSA Panel. Through scroll bars, the user can visualize the whole alignment, inspect the conservation histogram and the amino-acid distribution occurring at a fixed MSA

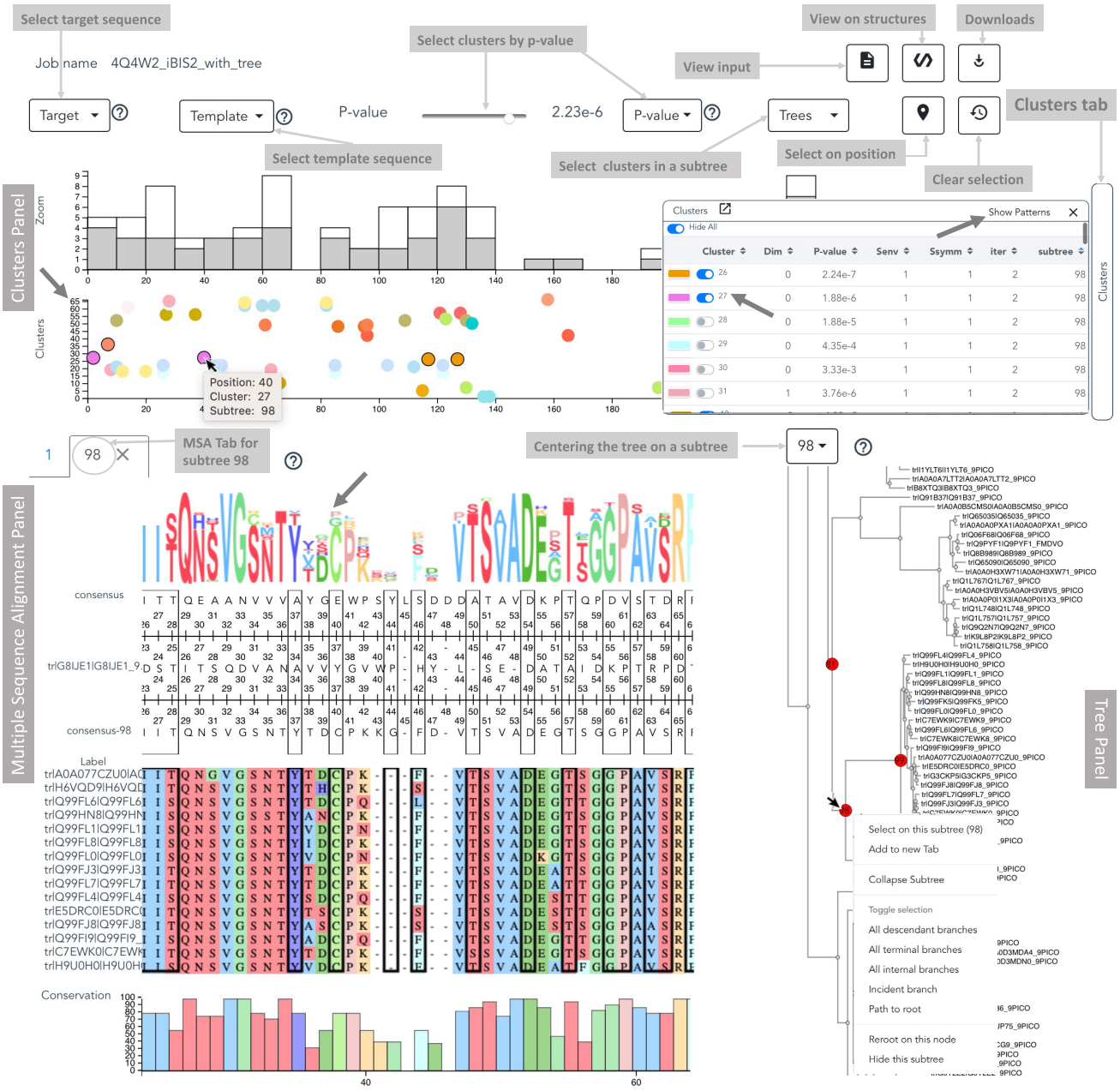


Figure 1. iBIS2Analyzer Results Page. Three main panels are dedicated to the Clusters, MSA and Tree analyses. The 'Clusters' tab grants access to the list of all clusters identified on the analysed subtrees (see pop-up table, top) and to their statistics (the button 'Show Patterns' opens a window with more details). The button 'View on structures' grants access to the Structural Visualization Page. A selection of clusters with *P*-value <2.23e−6 realized through the *P*-value bar appears on the Clusters Panel. The x-axes of the two plots 'Zoom' and 'Clusters' correspond to the MSA consensus sequence. In the 'Zoom' plot, the distribution of positions in the selected clusters is highlighted (grey) and compared to all coevolving positions (white). In the 'Clusters' plot, the selected clusters, out of 65 (y-axis), are displayed in colors; several clusters are identified by the undergoing analysis of subtree 98 and they are circled in black. Information on a position of a cluster can be accessed by hovering over the corresponding color dot (e.g. position 40 in cluster 27 identified in subtree 98). By clicking on node 98 in the tree structure, a pop-up panel opens and by clicking on 'Add to new Tab', the sub-MSA of subtree 98 is displayed in a dedicated tab. Position 40 is highlighted in the MSA because it belongs to the identified clusters. By scrolling the sub-MSA, the user can see the amino acids occurring at that position, compare them with the consensus sequence of the MSA of the full tree (showing a 'E', top of the graphical ruler), the consensus sequence of the MSA for subtree 98 (showing a 'C', bottom) or a specific sequence (showing a 'Y', center; chosen by the user with the 'Target' button—it can be the sequence of an associated structure, for instance), visualise the conservation level and the most frequent amino acid of the position in the sub-MSA. See the Documentation Page for further details.

position. Sequences in the MSA are colored by physico-chemical classes. A three layers graphical ruler describing the MSA consensus sequence is provided together with amino acid coordinates. It allows to map the coevolving positions to different sequences: the consensus sequence of the whole MSA (top layer), a sequence selected through the 'Target sequence' button (middle layer), a sequence selected through the 'Template sequence' button (bottom layer). For the whole MSA, the three layers are set with the MSA consensus sequence by default, while for a sub-MSA, the top layer is set with the MSA consensus sequence and the other two layers with the sub-MSA consensus sequence. See example in Figure 1.

The Tree Panel. Interactive panel where the phylogenetic tree can be moved manually and zoomed in and out. To facilitate visualization of large trees, the user can use the 'Centering on subtrees' button to automatically center the view on a specific subtree or to reset the view. All nodes in subtrees that are analyzed with iBIS^2 are colored red. By clicking on a red node (see Figure 1), several features can be selected to highlight branches, collapse subtrees, and mark paths from a node to the root. By selecting 'Select on this subtree', iBIS2Analyzer displays in the Clusters Plot all clusters identified in the sub-MSA of the subtree, and only those. This is a fast way to select clusters identified in a subtree. By selecting 'Add to new Tab', iBIS2Analyzer opens a new sub-MSA tab to explore the sequences in the subtree, as illustrated in Figure 1 for subtree 98. The user can select, one by one, several subtrees and open multiple tabs. Tree visualization is implemented by adapting the phylotree.js library (29).

Coordinated panels' interaction in the Results Page. Several dedicated buttons help to drive coevolution analysis and access the data.

The 'Target' button opens a menu where the user can specify a sequence that will be displayed in the graphical ruler of the MSA Panel. The user can choose any sequence in the MSA or upload a new sequence. The target sequence is aligned (30) against the 'Template' sequence which, by default, is the consensus sequence of the MSA. Note that new sequences will only be mapped to the MSA but will not be included in the MSA for analysis.

The 'Template' button opens a menu where the user can specify which sequence in the MSA will be used to align (via the Smith–Waterman algorithm (31)) the target sequence to the MSA. Through this selection, the user can control the alignment of its target sequence to the MSA by passing through a template sequence which is close to the target. If the target sequence is included in the MSA, it can be aligned with itself by choosing the template to be the target. It can also be aligned against any other sequence in the MSA and in this case, the alignment of the two sequences in the MSA is displayed. By default, the template sequence is the 'consensus' sequence.

The '*P*-value' bar helps the user select only clusters whose *P*-value is lower or equal to a certain threshold. See Figure 1 for an example, where only clusters with *P*-value $<2.23e{-}6$ are selected and displayed in the Clusters Panel. This option is particularly useful to filter the baseline correlations visualized in the Clusters' Plot.

The 'Trees' button opens a menu with the names of all subtrees analyzed by iBIS^2 and allows the user to display clusters identified for that subtree in the Clusters Panel and in the (sub-)MSA.

On the top right of the Results Page, several buttons are dedicated to multiple tasks: retrieving a job input, downloading all results, viewing clusters on structures. By clicking on the 'Select on Position' button, the user can choose to see, over the length of the MSA, only the clusters with at least one position in the zoomed region. The default view can be restored by clicking on the 'Clear Selection' button.

The Structural Visualization Page

Accessible through the 'View on Structures' button of the Results Page (Figure 1), the Structural Visualization Page is organized in several panels dedicated to the visualization of coevolving positions on protein structures (Figure 2). The page is organized with three panels which are borrowed from the Results page (Zoom plot, Clusters plot and Clusters Tab) and one or more other panels displaying protein structures uploaded by the user. For each structure, the user can provide either a PDB file, a CIF file or a PDB id, possibly containing multiple chains. Multiple structures can be loaded simultaneously side-by-side (Figure 2B). In the window opened for a PDB structure, chain-specific on/off switches allow each chain to be viewed or not. The user can enable/disable each cluster for visualization from the Clusters Tab. The same colors are used to identify clusters on the Clusters Panel and the structures (see Figure 2AB and DE). The possibility to upload multiple structures makes it a useful feature when either protein interactions or different foldings (e.g. disordered versus ordered regions) for the same protein are explored (Figure 2B). Note that residues in the structure are mapped on the MSA by retrieving the sequence of each PDB chain and aligning it on the MSA consensus sequence (30). Each chain sequence is mapped on the protein consensus sequence in a dedicated 'Projection' plot (Figure 2A).

Scores defining the confidence on a coevolution cluster

Three scores helping the user to evaluate the confidence in a cluster are provided: *P*-value, symmetric/Ssymm, environmental/Senv. Examples illustrating the three scores are given in the Documentation page and further details are found in (16,18). For the two last scores, see also (27).

The *P*-value score is computed with a Fisher test on a diagonal matrix, where the elements of the diagonal represent the co-evolution pattern satisfied by all positions in a cluster; for example, given a cluster of two positions in a MSA of 75 sequences, '24-30-21' is a pattern representing three distinct pairs of amino-acids on the two MSA positions that occur on subsets of 24, 30 and 21 sequences, respectively. The subsets are the same for the two positions and, in this case, we talk about a 'perfect pattern' of coevolution. When the pattern if not perfect, the *P*-value is computed on the maximum set of aligned sequences displaying a perfect pattern.

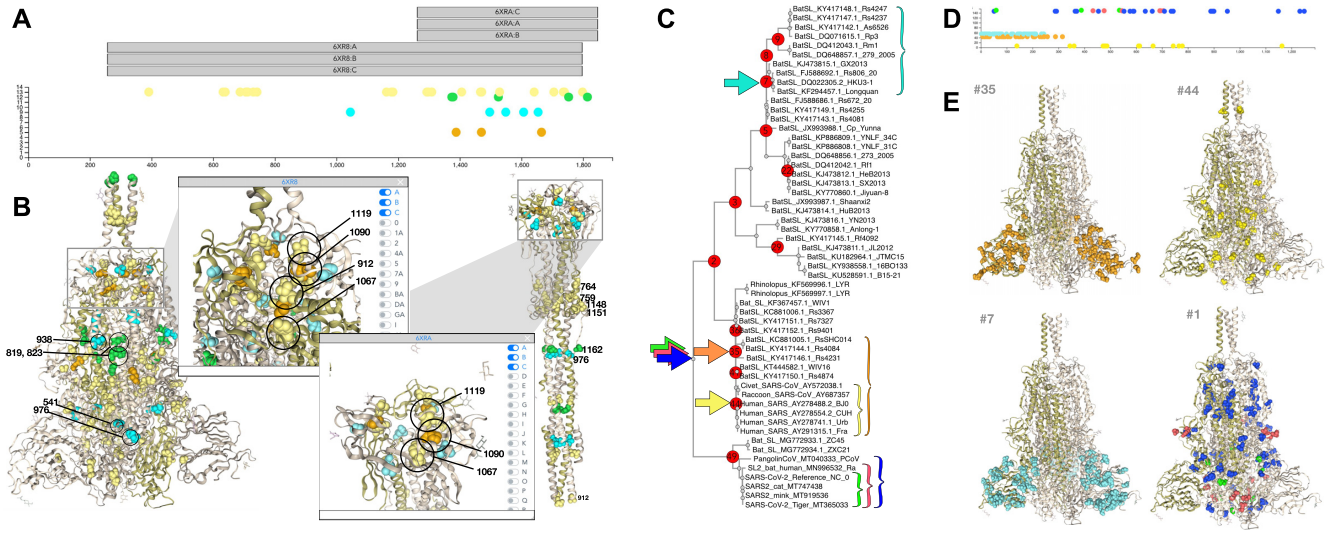


Figure 2. Two analyses of the Spike protein of the SARS-CoV-2 virus based on iBIS2Analyzer. (**A-B**) Coevolving residues of the Spike protein are plotted on the conformations associated to pre- and post-fusion. (**A**) The pre-fusion (6XR8:ABC) and post-fusion (6XRA:ABC) structures are mapped against the sequence. Four clusters are represented in the Clusters plot. (**B**) Residues in the four clusters (A) are plotted on the pre-fusion (left) and post-fusion (right) structures. Two panels zoom on structural details. (**C**) Tree of SARS sequences belonging to the betacoronavirus genera. (**D**) Clusters plot representing six clusters of coevolving residues. Cluster colors match arrow colors pointing at the root of subtrees in C where clusters were identified. (**E**) The six clusters in D are plotted on the pre-fusion structure.

The symmetric score *Ssymm* and the environmental score *Senv* are used in the CLAG clustering algorithm, designed to cluster the BIS^2 correlation matrix. They vary in the interval [0, 1] and they express the degree of 'similarity' of coevolution of positions in a cluster with respect to all other analysed positions. In particular, scores equal to 1 correspond to a cluster where all positions show an identical coevolution pattern with all other analysed positions. High scores guarantee the confidence in a cluster and because of this, iBIS2Analyzer outputs only clusters with both scores >0.5.

Downloads and implementation

iBIS2Analyzer provides a .zip archive containing all data issued by the analysis. The interface is implemented with NGL Viewer (http://nglviewer.org/#cite), a collection of tools for web-based molecular graphics, very fast and visualisable on smartphones (28).

RESULTS

Several BIS^2 (16–19) and iBIS^2 (20) coevolution analyses have been presented. They concern the detection of hotspot residues within a protein, fragments of residues in contact within a protein, correlations among residues in unfolded structures, long distance correlations between residues involved in conformational changes, analysis of viral polyproteins, secondary mutations in viruses emerged after drug treatments. In the *Tutorial page,* several examples are discussed. Here, we present two that highlight the importance of phylogenetically-driven coevolution analysis. iBIS^2 not only finds biological signals that may be different from direct contacts, but also allows us to profitably explore evolutionary events, like zoonosis, in novel and insightful ways.

Contacts in alternative conformations are found from independent subtrees analysis of the Spike protein

The Spike protein participates in the formation of the envelope of the SARS-CoV-2 virus and contains the binding receptor involved in the interaction with the host. We considered 25 sequences representing the alpha and beta coronavirus genera, spanning from bats to rodents and including human sequences. The beta coronavirus group is represented by four main lineages: the Embecovirus (containing the human sequences OC43 and HKU1), the Merbecovirus (MERS-CoV), the Nobecovirus (HKU9), and the Sarbecovirus (SARS-CoV and SARS-CoV-2). iBIS2Analyzer, run with the BIS^2 algorithm, highlights two clusters (yellow and orange in Figure 2AB) which separate the genera alpha and beta: they highlight sequence positions that are conserved within each lineage, while mutating across the two. Also, the orange cluster discriminates, with a different mutation, the Merbecovirus subgenus. A third cluster (cyan) separates the Sarbecovirus sequences from all other sequences and a fourth one (green) separates the Nobecovirus sequences from all others. Residues belonging to the same coevolution cluster can be far apart in one structure and in contact in the other: 759 and 1151 (yellow) are far apart in pre-fusion and in contact in post-fusion over different chains (AC, CB, BA). Residues in the same cluster can exchange their role: residue 1067 replaces in post-fusion the role of 912 in pre-fusion (yellow; inlays of Figure 2B). Different residues in cluster pairs can be in contact in different conformations: in pre-fusion, residue 938 (cyan) is in contact with 819 and 823 (green), whereas in post-fusion residue 976 (cyan) is paired with 1162 (green).

Finding zoonosis in coevolution signals of the Spike

'One time' multiple mutation events are good indicators of zoonosis in viruses. These events are detectable with iBIS[2]. For this, we realized a coevolution analysis of the exhaustive albeit limited set of Spike sequences which was previously considered in a seminal study on the evolution of SARS-CoV-2 (32). The phylogenetic tree is represented in Figure 2C and comprises sequences of the betacoronavirus genera. Subtree #1 shows two distinct branches separating SARS-CoV-1 and SARS-CoV-2 sequences: subtree #2 contains SARS1 and many human-unrelated sequences, and its sister tree contains human-hosted and human-infecting variants. Several observations can be put forward.

First, subtrees #7 (cyan) and #35 (orange) share many common coevolving pairs of residues localized on the N-terminal domain (NTD) of the protein. These coevolving residues are found on independent subtrees and this is an indication of their structural and/or functional importance.

Second, even though subtrees #35 (orange) and #44 (yellow) are one contained in the other, they highlight very different mutations. They gave origin to the SARS-CoV (SARS1) lineage. Those identified with the cyan subtree are associated to the 'animal' SARS1 and the ones identified by the yellow subtree are associated to SARS1 infecting 'humans'.

The three clusters, colored blue, red and green, associated to iBIS[2] analysis of subtree #1 identify coevolving residues whose amino acid changes occur on SARS2+SL2-bat+PangolinCoV (blue bracket), SARS2+SL2-bat (red bracket), SARS2 (green bracket) sequences. The 'blue' cluster shows a known recombination event characterized by a large number of mutations which are dispersed over the whole protein structure. The 'red' cluster shows coevolution between a mutation at the TMPRSS2 position and mutations on the Receptor Binding Domain (RBD). The 'green' cluster shows mutations that might affect the opening/closing RBD mechanism. Note that the three clusters are identified in subtree #49 but they persist in tree #1, showing the importance of the positions across the whole tree.

DISCUSSION

iBIS2Analyzer brings to light new rules of protein evolution, driven by functional and structural coevolution in a protein family. It conveys an automatic and highly customizable pipeline for phylogeny-driven coevolution analysis of protein sequences. It helps the user to explore coevolving positions possibly obtained for different subfamilies of a protein, of difficult access otherwise. A major advantage of iBIS2Analyzer design is that the user sees all relevant information on a single page, for analysis based on sequence and structure respectively, creating an interactive working environment that transfers his/her choices between panels. The Results and Structural Visualization pages have been designed to stimulate creative thinking on the problem, to foster hypotheses on protein behaviour and new strategies for the design of experiments.

FUNDING

Agence Nationale de Recherches sur le Sida et les Hepatites Virales [ANRS–AAP-2021-CSS-12]. Funding for open access charge: Agence Nationale de Recherches sur le Sida et les Hepatites Virales [ANRS–AAP-2021-CSS-12].
Conflict of interest statement. None declared.

REFERENCES

1. Lockless,S. and Ranganathan,R. (1999) Evolutionary conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
2. Suel,G., Lockless,S., Wall,M. and Ranganathan,R. (2003) Evolutionary conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, **23**, 59–69.
3. Baussand,J. and Carbone,A. (2009) A combinatorial approach to detect co-evolved amino acid networks in protein families with variable divergence. *PLoS Comput. Biol.*, **5**, e1000488.
4. Marks,D.S., Colwell,L. J., Sheridan,R., Hopf,T. A., Pagnani,A., Zecchina,R. and Sander,C. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, **6**, e28766.
5. Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1293–E1301.
6. Hopf,T.A., Colwell,L.J., Sheridan,R., Rost,B., Sander,C. and Marks,D.S. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.
7. Jones,D.T., Buchan,D.W.A., Cozzetto,D. and Pontil,M. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.
8. Morcos,F., Jana,B., Hwa,T. and Onuchic,J.N. (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 20533–20538.
9. Juan,D., Pazos,F. and Valencia,A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.
10. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
11. Tunyasuvunakool,K., Adler,J., Wu,Z., Green,T., Zielinski,M., Žídek,A., Bridgland,A., Cowie,A., Meyer,C., Laydon,A. *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, **596**, 590–596.
12. Hopf,T., Scharfe,C., Rodrigues,J., Green,A., Kohlbacher,O., Sander,C., Bonvin,A. and Marks,D. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, **3**, e03430.
13. Wang,S., Sun,S., Li,Z., Zhang,R. and Xu,J. (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, **13**, e1005324.
14. Kuriyan,J. (2004) Allostery and coupled sequence variation in nuclear hormone receptors. *Cell*, **116**, 354–356.
15. Del Sol,A., Fujihashi,H., Amoros,D. and Nussinov,R. (2006) Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.*, **2**, 2006.0019.
16. Dib,L. and Carbone,A. (2012) Protein fragments: functional and structural roles of their coevolution networks. *PLoS One*, **7**, e48124.
17. Champeimont,R., Laine,E., Hu,S.-W., Penin,F. and Carbone,A. (2016) Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. *Sci. Rep.*, **6**, 26401.
18. Oteri,F., Nadalin,F., Champeimont,R. and Carbone,A. (2017) BIS2Analyzer: a server for co-evolution analysis of conserved protein families. *Nucleic Acids Res.*, **45**, W307–W314.
19. Douam,F., Fusil,F., Enguehard,M., Dib,L., Nadalin,F., Schwaller,L., Hrebikova,G., Mancip,J., Mailly,L., Montserret,R. *et al.* (2018) A

protein coevolution method designed for conserved sequences uncovers critical features of the original HCV fusion mechanism and provides molecular basis for the design of effective antiviral strategies. *PLoS Pathogens*, **14**, e1006908.
20. Teppa,E., Nadalin,F., Combet,C., Zea,D.J., David,L. and Carbone,A. (2020) Coevolution analysis of amino-acids reveals diversified drug-resistance solutions in viral sequences: a case study of Hepatitis B virus. *Virus Evol.*, **6**, veaa006.
21. Guindon,S., Dufayard,J.-F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
22. Price,M.N., Dehal,P.S. and Arkin,A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, **26**, 1641–1650.
23. Gascuel,O. (1997) BIONJ, an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, **14**, 685–695.
24. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *CABIOS*, **8**, 275–282.
25. DOTREE, Plotree, DOTGRAM, Plotgram. (1989) PHYLIP-phylogeny inference package (version 3.2). *Cladistics*, **5**, 163–166.
26. Gouy,M., Guindon,S. and Gascuel,O. (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221–224.
27. Dib,L. and Carbone,A. (2012) CLAG, an unsupervised non hierarchical clustering algorithm handling biological data. *BMC Bioinformatics*, **13**, 194.
28. Rose,A.S., Bradley,A.R., Valasatava,Y., Duarte,J.M., Prlić,A. and Rose,P.W. (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
29. Shank,S.D., Weaver,S. and Pond,S.L.K. (2018) phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics*, **19**, 276.
30. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
31. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
32. Hoffmann,M., Kleine-Weber,H., Schroeder,S., Krüger,N., Herrler,T., Erichsen,S., Schiergens,T.S., Herrler,G., Wu,N.H., Nitsche,A. *et al.* (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, **181**, 271–280.