Democritus University of Thrace
School of Health Sciences
Department of Molecular biology and Genetics

Conserved clusters of contacts in protein structures

DIPLOMA DISSERTATION

George Kolypetris
Computational and structural biology laboratory, NMG group
Department of Molecular biology and Genetics, Democritus University of Thrace

Dr. Nicholas M. Glykos, Associate professor
Computational and structural biology laboratory, NMG group
Department of Molecular biology and Genetics, Democritus University of Thrace

Alexandroupolis, Evros | June 2022

Δημοκρίτειο Πανεπιστήμιο Θράκης
Σχολή Επιστημών Υγείας
Τμήμα Μοριακής βιολογίας και Γενετικής

Συντηρημένες συστοιχίες επαφών σε πρωτεϊνικές δομές

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γεώργιος Κολυπέτρης
Εργαστήριο υπολογιστικής και δομικής βιολογίας, ομάδα ΝΜΓ
Τμήμα Μοριακής βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Δρ. Νικόλαος Μ. Γλυκός, Αναπληρωτής καθηγητής
Εργαστήριο υπολογιστικής και δομικής βιολογίας, ομάδα ΝΜΓ
Τμήμα Μοριακής βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Αλεξανδρούπολη, Έβρος | Ιούνιος 2022

acknowledgements

I would like to thank Dr. Nicholas M. Glykos for allowing me to undertake my diploma dissertation in his research group. It was an enlightening experience that taught me much about how to be a scientist.

I would also like to thank my colleagues at the NMG group and my friends at Democritus Industrial Robotics (DIR), for all the fun time we had together.

Lastly, I would like to thank my family, for their constant support during my undergraduate studies at the MBG department. I would not be the man I am today without them.

And even if you cannot make your life the way you
want it,
this much, at least, try to do
as much as you can: don't cheapen it
with too much intercourse with society,
with too much movement and conversation.

Don't cheapen it by taking it about,
making the rounds with it, exposing it
to the everyday inanity
of relations and connections,
so it becomes like a stranger, burdensome.

~ Constantine P. Cavafy, 1913

table of contents

abstract

Proteins are able to adopt various structures and functions due to interactions among amino acids in atomic level. These interactions are affected both by residual preference, as well as extrinsic effectors. To study these interactions, we performed a 5Å weighted, all-atom analysis, prioritising hydrophobic interactions and Coulomb forces, as well as a 7Å unweighted C_α analysis, on protein structures categorised by taxonomy classification and temperature. In terms of taxonomy, organisms were divided into their respective domains, while in terms of temperature, organisms were classified according to their optimum growth temperature. Results from both analyses highlighted the existence of conserved clusters of contacts in globular protein structures, as well as distinct differences in contact patterns in every group of study. Lastly, whilst the methodology developed still requires further analyses in evolution groups and other environmental factors, it provides a supplementary method to be performed alongside other available techniques for protein or proteome analysis.

protein structures contacts amino acid networks amino acid interactions AAN Protein Data Bank PDB PISCES PERL R bacteria archaea eukarya mesophiles thermophiles optimum growth temperature domains

περίληψη

Οι πρωτεΐνες μπορούν να υιοθετούν διάφορες δομές και λειτουργίες λόγω των αλληλεπιδράσεων μεταξύ των αμινοξέων του σε ατομικό επίπεδο. Οι αλληλεπιδράσεις αυτές επηρεάζονται τόσο από τις προτιμήσεις των καταλοίπων, όσο και από εξωτερικούς παράγοντες. Για να μελετήσουμε αυτές τις αλληλεπιδράσεις, πραγματοποιήσαμε μια κατευθυνόμενη, πλήρης ατόμων ανάλυση 5Å, δίνοντας προτεραιότητα σε υδροφοβικές αλληλεπιδράσεις και δυνάμεις Coulomb, όπως και μια μη κατευθυνόμενη, C_α ανάλυση 7Å, σε πρωτεϊνικές δομές κατηγοριοποιημένες βάσει ταξονομικής κατάταξης και θερμοκρασίας. Ταξονομικά, οι οργανισμοί κατηγοριοποιήθηκαν στις αντίστοιχες επικράτειες, ενώ από άποψη θερμοκρασίας, οι οργανισμοί κατατάχθηκαν σύμφωνα με τη βέλτιστη θερμοκρασία ανάπτυξής τους. Τα αποτελέσματα των αναλύσεων ανέδειξαν την ύπαρξη συντηρημένων συστοιχιών επαφών σε πρωτεϊνικές δομές, όπως και διακριτές διαφορές σε κάθε ομάδα μελέτης. Τέλος, ενώ η ανεπτυγμένη μεθοδολογία απαιτεί περαιτέρω αναλύσεις σε εξελικτικές ομάδες και άλλους περιβαλλοντικούς παράγοντες, παρέχει μια συμπληρωματική μέθοδο που θα μπορούσε να εφαρμοστεί μαζί με άλλες διαθέσιμες τεχνικές για ανάλυση πρωτεϊνών ή πρωτεωμάτων.

Πρωτεϊνικές δομές επαφές αμινοξικά δίκτυα αμινοξικές αλληλεπιδράσεις AAN Πρωτεϊνική Βάση Δεδομένων PDB PISCES PERL R βακτήρια αρχαία ευκάρυα μεσόφιλα θερμόφιλα βέλτιστη θερμοκρασία ανάπτυξης επικράτειες

1.1 Protein structure

Proteins are one of the four major macromolecules that direct life (Stollar and Smith, 2020). They participate in almost every biological process, as catalysts, transport and storage media, mechanical support, movement factors or in many more other functions (Berg et *al*, 2015; Nelson and Cox, 2017). This versatility in both form and function is achieved through their assembly by a set of 20 amino acids, each with their own unique properties, as well as their hierarchical levels of structure.

Proteins are structured in four levels, with each level being directly dependent of their previous ones and contributing significantly to the protein's final structure and function. First, the primary structure is defined as the sequence of the amino acid residues. Experiments conducted by Anfinsen in the 70's showed that the amino acid sequence alone is definitive of the protein's final, three-dimensional structure. The amino acid sequence also includes the location of disulphide bonds that covalently link different parts of the polypeptide chain together (Caetanno-Anollés et *al.*, 2009; Stollar and Smith, 2020).

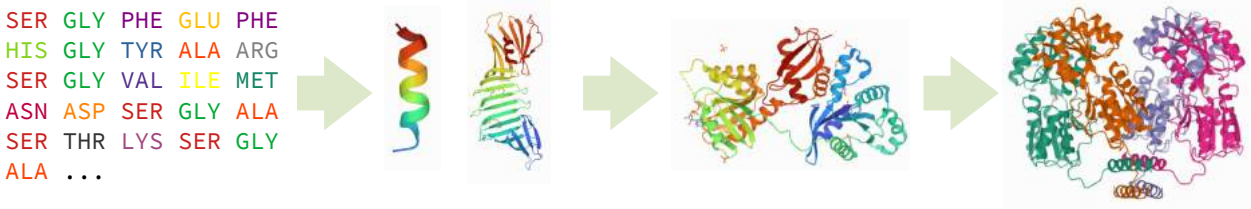


Illustration 1 Levels of protein structure

Proteins are structured in four levels. In the primary structure, the sequence of the amino acid residues will dictate the final, three-dimensional structure of the protein. In the secondary structure, the neighbouring residues will begin to interact through non-covalent, hydrogen bonds to form α-helices and β-sheets. These structures will be combined to form super-secondary complexes, which would lead to the formation of domains in the tertiary structure. In this level, proteins acquire their biological functionality and three-dimensional structure, through both covalent (e.g. disulphide bonds) and non-covalent interactions (hydrogen bonds, salt bridges, hydrophobic interactions). The quaternary level is reserved for multi-polyptide chain proteins, using all kinds of bonds mentioned to retain its form and function.

In regards to the amino acids themselves, it is important to review some of their properties. Every amino acid is comprised of a backbone group, consisting of a central α carbon atom (C_α) that is joined with an amino group ($-NH_2$), a carboxyl group (-COOH) and a hydrogen atom, as well as a side chain, which is also joined with the C_α atom. This side chain is unique for every amino acid and provides many of their distinct features. Despite their uniqueness though, these side chains share common chemical and physical properties, which allow us to group their respective amino acids. The most common way of grouping the 20 amino acids is by their residual polarity. As such, there are hydrophobic amino acids (alanine (A), valine (V), leucine (L), isoleucine (I), phenylalanine (F), proline (P), methionine (M) and glycine (G)), polar amino acids (serine (S), threonine (T), cysteine (C), asparagine (N), glutamine (Q), histidine (H), tyrosine (Y) and tryptophan (W)) and charged residues (aspartic acid (D), glutamic acid (E), lysine (K) and arginine (R)). An illustration of all 20 amino acids, grouped by residual polarity, can be found in Figure 1.

Amino acids can be arranged in any order in the polypeptide chain, linked through a peptide bond, which formulates between the carboxyl group of the first amino acid with the amino group of the second amino acid. This bond, due to conjugation in the carboxyl group, is partially double, therefore making it flat in the three-dimensional space. Whilst this restrains the amino and carboxyl groups, it does not restrain the central α carbon atom, which can adopt

various angles around these groups. The flexibility of the C_α atom is crucial for the next level of protein structure, the secondary structure.

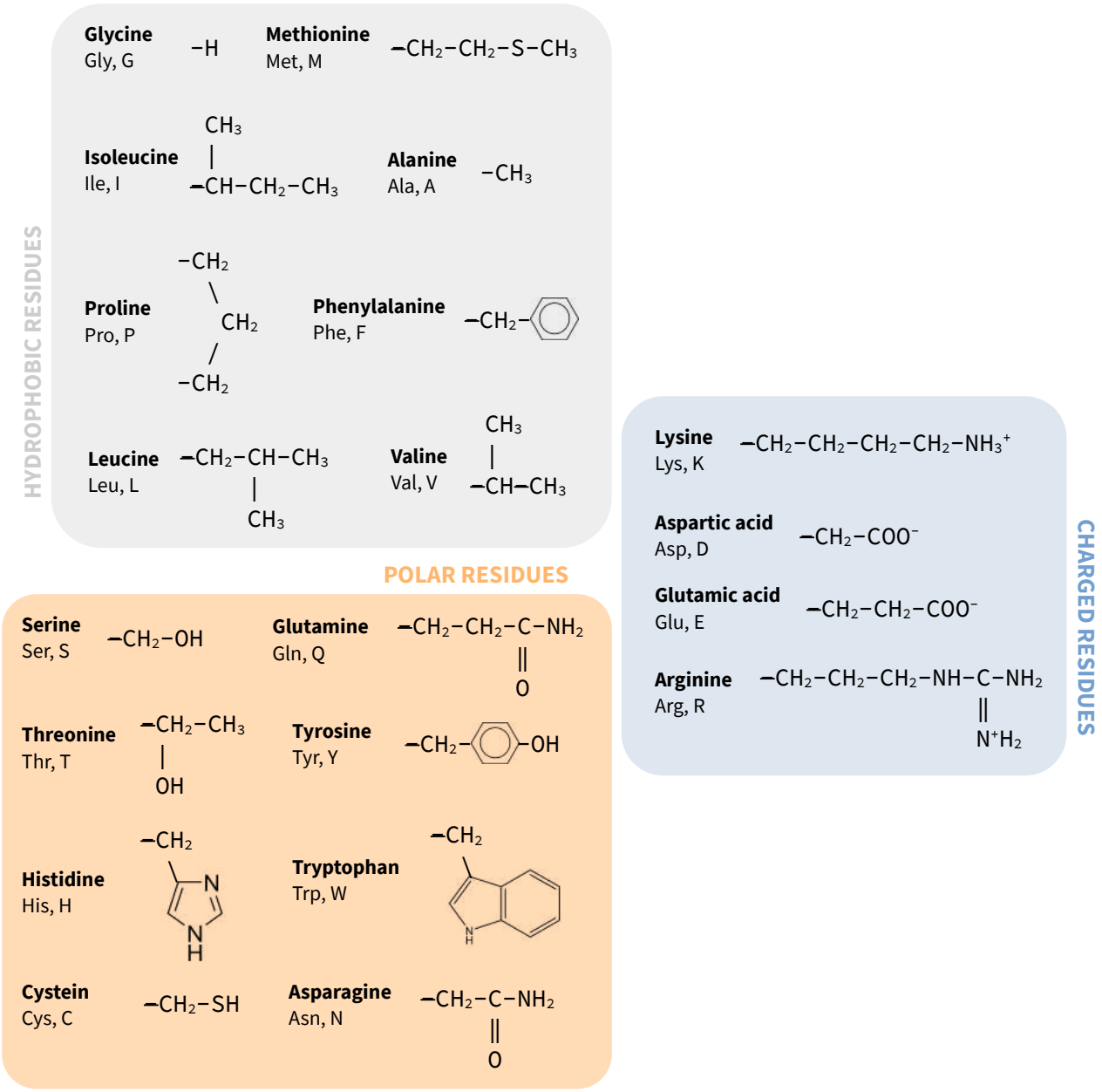


Figure 1 Amino acids grouped by residual polarity

There are 20 amino acids, each with a different side chain, which provide many of their properties. However, the side chains share both physical and chemical properties, allowing us to group the respective amino acids. The most popular way of classifying the amino acid residues is by their residual polarity. As such, amino acids are divided into hydrophobic, polar and charged. Specific atoms in each residual chain are responsible for this classification. All carbon atoms provide potential hydrophobic interactions, whereas oxygen, nitrogen and sulphur (in cysteines only) provide electrostatic interactions and hydrogen bonds, due to their respective electronegativity or electropositivity.

The secondary structure is the level of protein organisation where non-covalent bonds, hydrogen bonds to be precise, start affecting the protein's three-dimensional structure. The amino acids, according to their position in the polypeptide chain, begin forming basic, energetically favourable structures, in cooperation with their neighbouring residues. This synergy results in the formation of structures like α-helices and β-sheets, which are held together ex-

clusively by hydrogen bonds (Herschlag and Pinney, 2018; Trent Kemp et *al.*, 2021). These structures are usually abundant in the hydrophobic core of the protein and the hydrogen bonds formed are among the backbone atoms of each amino acid, although there have been occasions of side chain atoms being part of secondary structures.

Having formulated basic structures, the protein proceeds with organising its components even further, by combining secondary elements to create formations called super-secondary structures or motifs. These formations either perform specific functions, like DNA binding, or are part of larger, more complex structural or functional systems, like domains (Branden and Tooze, 1999).

Domains are distinct areas of a protein that can fold independently into a stable structure with a dedicated function or a set of functions. They are a basic component of the third level of protein organisation, the tertiary structure. In this level, proteins acquire their biological activity and utilise all available bonds, both covalent and non-covalent, to form and retain their biologically functional shape. Such bonds are the hydrogen bonds that we mentioned earlier, as well as disulphide bonds (created by conserved cysteines), salt bridges (which are formed by oppositely charged residues) (Bosshard et *al.*, 2004; Bandyopadhyay et *al.*, 2020) and hydrophobic interactions, which stabilise the protein core and the protein's shape in a three-dimensional space (Raschke, 2006; Mirzaie, 2017). It is also worth noting that some tertiary structures require the assistance of metals or even water molecules to retain their functionality (Branden and Tooze, 1999).

The quaternary structure is the last level of protein organisation. It is reserved for proteins consisting of more than one polypeptide chain, either identical or different. As with tertiary structure, both covalent and non-covalent bonds are at force, for both formation and maintenance of the three-dimensional structure of the complex.

1.2 Amino acid networks

Proteins have been widely studied based on their structural analysis: following the 4-level structure, we have made observations about protein structures as a group of secondary and super-secondary structures and their respective interactions. Over the past decades though, the advancement of computational tools have given rise to a new approach, one that utilises the tools of network analysis (Csermely, 2008). According to the network characterisation technique, protein three-dimensional structures are transformed into an amino acid network, where we study the interactions between amino acids as a whole and not as part of secondary or super-secondary interactions (Khor, 2012; Yan et *al.*, 2014).

Following the principles of network analysis, nodes denote the system elements and edges denote their interactions. These interactions can be weighted to characterise the edge's strength (Greene, 2012; Yan et *al.*, 2014). For our research, nodes are considered the amino acids of the protein and their interactions are the network's edges. These networks have multiple names in available references, their most common being amino acid networks (AAN) (D'Amico et *al.*, 2020; Viswanathan et *al.*, 2015), protein structure networks (PSN)(Csermely, 2008; Greene, 2012) or residue interaction networks (RIP)(Hu et *al.*, 2014). For our research, we use the term amino acid networks (AAN).

The construction of an amino acid network requires the definition of several parameters. Firstly, attention needs to be given on whether our network will be weighted or not. A weighted network will treat specific interactions as more important than others, whilst an unweighted network will treat all interactions as equal (Khor, 2012; Karain and Quaraeen, 2017). Most amino acid networks are unweighted, with the edges being the physical distances between the nodes. Secondly, we need to define both our nodes and our edges. There are two main approaches in regards to node selection: we can either study the interactions of the C_α or C_β of each amino acid, or we can study the interactions of all atoms of each amino acid. While the second approach is obviously more accurate, it is computationally intensive (Greene, 2012). Table 1 includes a list of the most popular methods for the construction of an AAN.

Node selection directly affects edge selection as well. A C_α study will require a physical distance cutoff of 6.8-7 Angstrom for our edges, while an all atom study will require a physical distance cutoff of 4.5-5 Angstrom, the maximum distance between two atoms without a water molecule intervening in their interaction (Hu et *al.*, 2014; Viswanathan et *al.*, 2015). Lastly, it is important to remember that in proteins, the overall three-dimensional structure is maintained by both covalent and non-covalent bonds. As such, it would be inaccurate of an amino acid network not to consider the importance of the physical and chemical properties of the amino acids. Therefore, AANs also take into consideration Coulomb forces (electrostatic interaction energy), as well as van der Waals interaction energy.

Table 1 Methods for amino acid network construction (reproduced without permission from Yan et *al.*, 2014)

Nodes	Links	Network type
C_α	Node distance less than a threshold 7Å, 8Å and 8.5Å	Unweighted
C_β (C_α for Gly)	Node distance less than a threshold of 7Å and 8.5Å	Unweighted
Centroids of side chain	Node distance less than R_c=8.5Å	Unweighted
Amino acid	If distance between any atoms from the whole amino acids or only from the side chain is less than R_c=5Å Strength of non-covalent interactions based on atom-atom contact and only consider atom from side chain. Atom-atom distance cutoff: R_c=5Å	Unweighted
Amino acid	Links: if distance between any atoms from the whole amino acid residue or only from the side chain is less than R_c=5Å Weight: The number of possible atom-atom links	Weighted
Atom	The summation of the electrostatic interaction energy (Coulomb potential) and the van der Waals interaction energy (Lennard-Jones potential) between two atoms	Weighted
Amino acid	The summation of the electrostatic interaction energy (Coulomb potential) and the van der Waals interaction energy (Lennard-Jones potential) between two amino acid	Weighted
Geometrical centre of the side chain	Links: Node distance less than a threshold Weight: Miyazawa and Jernigan contact energy between two residues	Weighted
Amino acid	Links: within a cutoff distance or for at least 75% of an MD Weight: based on cross-correlation between the monomer over the course of the MD simulation	Weighted

1.3 Temperature as an extrinsic effector

While it is useful to study the internal structure of proteins, it is also important to keep in mind that their structure is directly affected by the environment in which they are produced, as well as the one they perform their designated function. There is a large number of extrinsic effectors on a protein's structure, notably temperature, pH, salinity or hydrostatic pressure. Each have their influence on a protein's composition, as well as its three-dimensional structure. For example, halophiles shift their polar and non-polar amino acids ratio to cope with high salt concentrations, while acidophiles and alkalophiles adapt only their peripheral structures (Jaenicke, 1991).

However, the main focus of the study of extrinsic effectors has always been over temperature, due to the large number of isolated organisms, as well as their potential usefulness in the biotechnological industry (Finch and Kim, 2018; Sterner and Liebl, 2001). The temperature spectrum of life has been found to range from -5°C to 110°C, with organisms being organised by their optimum growth temperature (Jaenicke, 1991; Finch and Kim, 2018). As such, there are psychrophiles, found in environments with temperatures from -5°C to 15°C, mesophiles, living in habitats with an average temperature of 15°C to 45°C, thermophiles, indigenous to habitats of temperatures from 45°C to 85°C and hyperthermophiles, which live on the very limit of viability, in temperatures from 85°C to 110°C. Given the minute number of psychrophiles isolated, our focus has been shifted to the study of mesophiles, thermophiles and hyperthermophiles.

Mesophiles constitute the majority of organisms isolated, given the favourable environmental conditions on which they thrive. Temperatures ranging from 15 to 45 degrees Celsius can be found on most of the biosphere, thus promoting the proliferation and dominance of mesophiles all over the planet. Thermophiles and hyperthermophiles on the other hand, can be found only in extreme environments, like hot springs or hydrothermal vents. These extreme habitats force these organisms to adapt to the environmental conditions, therefore providing us with valuable insight on the mechanisms that sustain life.

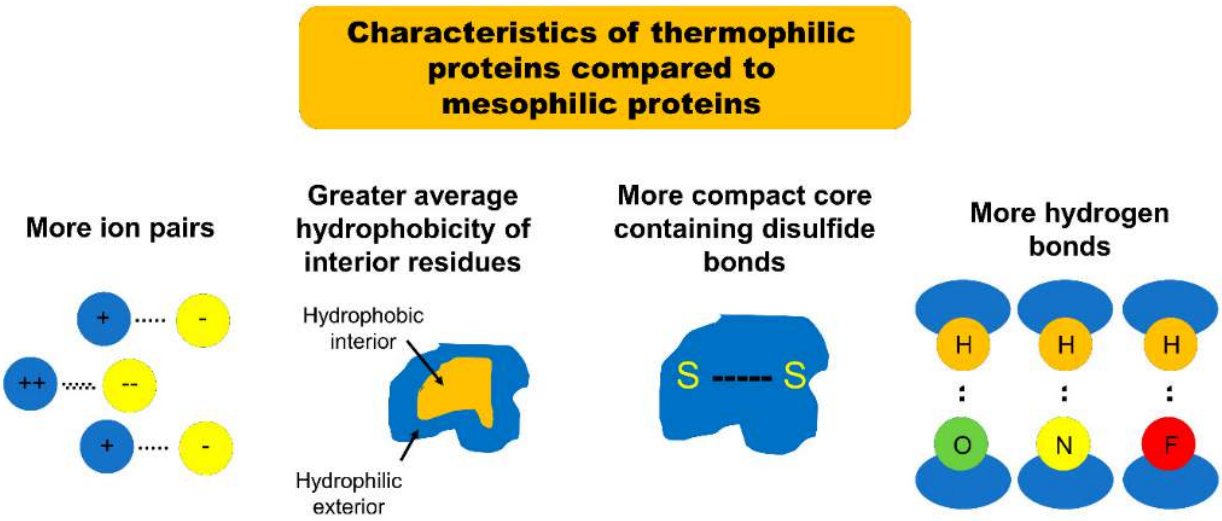


Figure 2 Characteristics of thermophilic proteins when compared to mesophilic proteins (reproduced without permission from Finch and Kim, 2018)

In regards to their protein structure, thermophilic and hyperthermophilic proteins have some distinct differences from their mesophilic counterparts. A graphic summary of these differences can be seen in Figure 2. As many studies have reported over the years, thermophilic proteins project a larger number of ion pairs that stabilise the three-dimensional structure. Specifically, thermophilic proteins show an increase in charged residues, especially Glu, Arg and Lys, while at the same time, there is a decrease of uncharged polar residues, like Gln, Asn, Thr and Ser (Sterner

and Leibl, 2008; Finch and Kim, 2018). Thermophilic proteins also display an increased number of buried salt bridges (Sterner and Liebl, 2008). These ion pairs were found to be strategically placed on the protein's amino acid sequence, as well as on the corresponding nucleotide sequence. Furthermore, these ion pairs have shorter distances when in comparison with their mesophilic counterparts (4 Angstrom for thermophilic ion pairs, when the average distance of mesophilic ion pairs in 6-8 Angstrom), thus making these pairs stronger (Szilagyi and Zavodszky, 2000; Sterner and Liebl, 2008).

Another difference among thermophilic and mesophilic proteins is the average hydrophobicity of the amino acid chains buried within the protein (Finch and Kim, 2018). Thermophilic proteins have a core even more hydrophobic than the mesophilic one. This increased hydrophobicity can be achieved through increased α-helical content and α-helix stability, increased compactness or packing densities, replacement of residues with energetically unfavourable conformations by glycine and optimised hydrophobic interactions (Kumar and Nussinov, 2001; Sterner and Liebl, 2008). For example, thermophilic helices showed a decreased number of β-branched amino acids, like Val, Ile and Thr, leading to more stable α-helices in thermophilic proteins (Sterner and Liebl, 2008). However, increased hydrophobicity is only observed in the protein core. Both the thermophilic and mesophilic proteins have the same hydrophobicity in regards to their solvent accessible surfaces (Finch and Kim, 2018).

Lastly, thermophilic proteins have more hydrogen bonds than their mesophilic counterparts. Thermophiles use hydrogen bonds to connect residues both within the core of the protein as well as on the exterior (Kumar and Nussinov, 2001). Hydrogen bonds were also used to bridge inner and outer amino acids and therefore stabilise the protein even more, despite the fact that hydrogen bonds become weaker at higher temperatures (Jaenicke, 1991; Finch and Kim, 2018).

1.4 Our goal

Protein structures are the result of amino acid interactions in the three-dimensional space, via definite levels of organisation. Despite their shared process of formation though, proteins exhibit a wide range of amino acid composition, structure, shape and function. Furthermore, these features are directly affected by extrinsic effectors, like temperature and others. This lead us to the question: How proteins retain their morphological, functional or compositional diversity, whilst following the same basic principles of formation?

One of our hypotheses has been the possible existence of conserved patterns of amino acid interactions. One should not confuse amino acid interactions in protein structures with amino acid composition of proteins. Amino acid interactions refer to the network of interactions among amino acids, regardless of their overall number, while amino acid composition studies the number of each amino acid in proteins, regardless of their interactions. Given the wide compositional diversity we mentioned previously, it would not make sense to use a variable parameter, when studying a universal process for proteins.

If conserved amino acid interactions were to exist, they would be able to provide a guide through the formation process of proteins, as well as support the multifaceted diversity we mentioned earlier. To this end, we developed a methodology for analysing protein structures, on their amino acid interaction level and not as α-helices or β-sheets. This approach can provide insight to the possible preference of interactions among amino acids, as well as highlight conserved patterns throughout evolution. Lastly, we also considered the effect of environmental factors and performed relevant studies.

2.1 Amino acid network

Amino acid networks are a widely used tool for computational analysis of proteins. The traditional protein structure of α-helices and β-sheets is transformed into a complex network of interactions among amino acids. This network can include all atoms of an amino acid, or just the C_α/C_β atoms of each amino acid, while consideration can be taken for physical and chemical properties of the protein's residues.

For our research, we have conducted our analysis on both the C_α level and the all-atoms level. For the C_α level, we have defined C_α of each amino acid as our node and a physical distance cutoff of 7 Angstrom as our edge. In this level, our network is unweighted, considering all interactions equal. For the all-atoms level, each atom of every amino acid is a node, but at the same time, it belongs to the larger, amino acid node complex. Interactions between atoms belonging to neighbouring amino acids are not acceptable, with our edge being defined as the physical distance of maximum 5 Angstrom. In this level, our network is weighted, accepting only electrostatically favourable interactions, as well as all the hydrophobic ones.

2.2 Protein Data Bank (PDB)

The Protein Data Bank is one of the world's largest databases for three-dimensional, biomolecular structures. It contains structures of proteins, nucleic acids as well as protein – nucleic acid complexes, which have been obtained through a number of structural biology methods, notably X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and cryo-electron microscopy (cryo-EM) (Berman et *al.*, 2000; Bahzadi and Gajdács, 2021). The worldwide database is a major source of information for relevant fields of research, like computational and structural biology, structural genomics, as well as proteomics.

Access to the PDB structures is free via the websites of its member organisations (PDBe, PRBj, RCSB and BMRB). For our research, we downloaded a number of structures from PDB's worldwide ftp server (https://ftp.wwpdb.org/), using the download.pl script (full script available in the Appendix). A total of 21.546 structures were successfully downloaded.

2.3 PISCES

PISCES is a protein sequence culling server (http://dunbrack.fccc.edu/pisces/), developed by Dunbrack Roland and Wang Guoli (Wang and Dunbrack, 2003). Based on user provided criteria, PISCES searches the Protein Data Bank "to provide the longest lists possible of the highest resolution structures that fulfil the sequence identity and structural quality cutoffs". Ultimately, PISCES provides lists of non-redundant entries from the entire Protein Data Bank, that is, a list of entries not including duplicates. The criteria available include minimum and maximum resolution of structure, minimum and maximum chain length, maximum pairwise percent sequence identity, inclusion of structures with chain breaks, NMR spectroscopy entries and X-ray crystallography entries. When X-ray entries are selected, maximum R-value is another available criterium.

For our research, we selected structures with 50% maximum pairwise percent sequence identity, a minimum resolution of 0.0 Å, a maximum resolution of 3.0 Å, X-ray crystallography entries only, a maximum R-value of 0.3 and chains with no breaks but with missing residues due to disorder. A list of 22.759 entries was provided.

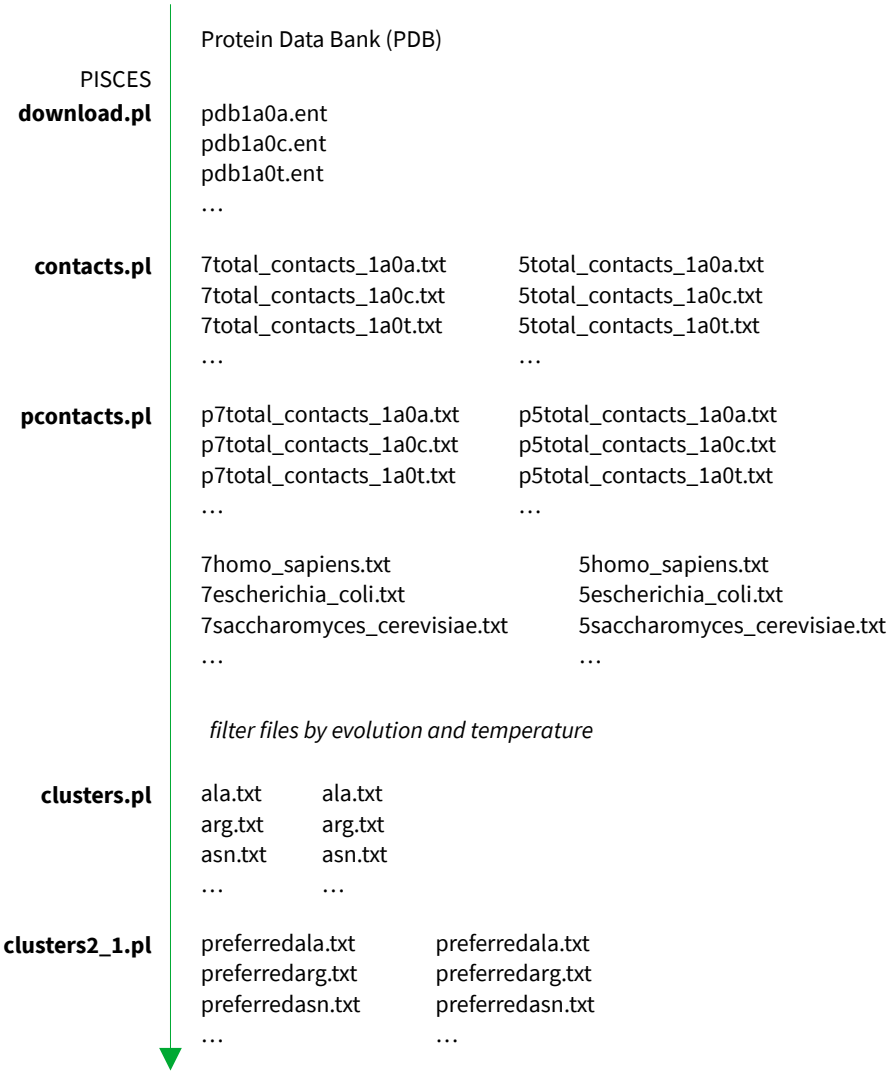


Illustration 2 Data acquisition and analysis

For our research, we downloaded, using the list provided by PISCES and the download.pl script, several protein structures from the Protein Data Bank's worldwide server (https://ftp.wwpdb.org/). These structures (pdb1a0a.ent etc.) were then processed by contacts.pl, to calculate the summary of all contacts between atoms of the structure. Contacts were calculated in two stages, on a C_α level (with the distance cutoff equal to 7 Angstrom) and on an atomic level (with a distance cutoff of 5 Angstrom). The generated files (7total_contacts_1a0a.txt etc and 5total_contacts_1a0a.txt etc) were processed by pcontacts.pl, to filter contacts based on electrostatic charge and hydrophobic interactions. The remaining contacts (p7total_contacts_1a0a.txt etc and 5total_contacts_1a0a.txt etc) were grouped together in species, generating the species files, both in 7Å and 5Å (hence 7homo_sapiens.txt etc and 5homo_sapiens.txt etc). Since our analyses focused on evolutionary relations and temperature for each analysis, we grouped the species files according to the parameter studied (e.g. for temperature, we grouped species in psychrophiles, mesophiles, thermophiles and hyperthermophiles). Lastly, using the clusters.pl and clusters2_1.pl scripts, we calculated the percentage of amino acid interaction for each amino acid. These data were later visualised by plots.r (not shown).

2.4 PERL programming language

The Practical Extraction and Reporting Language, also known as PERL, is a versatile, open-source programming language. PERL is extremely practical when manipulating alphanumeric sequences of characters, through the use of regular expressions and is capable of object-oriented programming. PERL is also a cross-platform language, supported on almost all available operating systems.

All of these reasons made PERL a natural choice for our programming language of use over the course of our research. Almost all of our scripts have been written using PERL, with the exception of plots.r, our data visualisation script. In detail, we scripted 5 main programs for conducting our analyses: download.pl, which establishes a connection with PDB's worldwide ftp server and downloads all the structures provided by PISCES, contacts.pl, which calculates all the contacts among the amino acids and pcontacts.pl, which filters the data provided by contacts.pl, based on chemical preferences of amino acids, electrostatic charges and hydrophobic interactions, in particular. Lastly, clusters.pl and clusters2_1.pl calculate the percentage of interactions among amino acids, providing a logarithmic percentage interaction table for plots.r to visualise. A complete description of our scripts and their content can be found in the Appendix.

2.5 R programming language

R is an easy and versatile programming language, focusing on statistical analysis and graphic presentation. It is also open-source and supports object-oriented programming facilities (Νικολάου, 2019). Lastly, R is a cross-platform language, supported on Windows, MacOS X and various Linux distributions.

For our research, we used Yan Holtz's algorithm and called it plots.r. The purpose of this script is to visualise our data in a circular treemap. Both the script and its description are available in the Appendix.

3.1 Amino acid interactions in bacteria, archaea and eukarya

Study of all structures based on taxonomy has provided interesting results regarding the interaction preferences of the amino acids both in general and for each domain individually. The structures were classified by their respective organism's domain (bacteria, archaea and eukarya) and then processed by clusters.pl and clusters2_1.pl. Logarithmic interaction percentages were arbitrarily categorised in three classes: high interaction percentages for values greater or equal to 1.86 (26%), medium interaction percentage for values greater or equal to 1.55 (20%) and minute interaction percentages for values lower than 1.55 (20%).

Overall, leucine has been highlighted as a high interaction percentage preference for all amino acids, both in 5Å and 7Å analysis. Isoleucine has been found to be a medium interaction percentage preference, whilst alanine, glycine and valine show a mixed interaction percentage preference pattern, depending on the domain of the organism, as well as the amino acid studied. In terms of polarity and charge, aspartic and glutamic acid seem to be a universal interaction percentage preference for polar and charged amino acids, while cystein has been found to interact mostly with hydrophobic residues and itself, with the exception of bacteria. Further preferences are discussed in detail in the following sections. A complete list of all interaction percentage preferences, both in tables and illustrations, is available in the Appendix.

3.1.1 Bacteria

In bacteria, all amino acids showed both high and medium interaction percentages with alanine, leucine and valine. Specifically, the 7Å analysis showed that all residues interact highly with leucine and alanine, when valine has both high and medium level interactions, depending on the residue in question. The 5Å analysis showed constantly high interactions only for leucine, with alanine and valine displaying both high and medium interaction percentage patterns. Surprisingly, a mainly high interaction percentage pattern was observed for glycine, but only in the 7Å analysis, since the pattern reduces to a medium one with all amino acids in its 5Å counterpart.

On the other hand, alanine and valine themselves have a rather small number of notable interactions, when in comparison with their interactions with all residues. In the 7Å analysis, alanine and valine interact highly with alanine, glycine, leucine and valine, followed by medium interactions with aspartic acid, serine and threonine. In the 5Å analysis, the aforementioned percentages remain mostly similar, with the exception of glycine, being reduced to a medium interaction percentage. Leucine is involved in the formation of the greatest number of contacts, mainly with alanine, glycine, isoleucine, leucine and valine in the 7Å analysis and alanine, isoleucine, leucine, phenylalanine and valine in the 5Å analysis. Interestingly, isoleucine shares similar interaction percentages with leucine, but it does not share leucine's interactivity with all residues. Lastly, glycine interacts highly with alanine, glycine, leucine and valine, followed by medium interactions with arginine, glutamic and aspartic acid, isoleucine, serine and threonine in both analyses.

In terms of polar and charged amino acids, a large number of interaction percentages, both high and medium, are observed for aspartic acid, serine and threonine, in both analyses. In detail, arginine, glutamine, glutamic acid, serine and threonine interact highly with aspartic acid, followed by medium interactions with alanine, aspartic acid and proline. Aspartic acid, asparagine, proline, cystein and tyrosine have high interaction percentages for serine, followed by medium interaction percentages from glutamine, glutamic acid in both analyses. Similar percentages were accounted for threonine. Figure 3 and Figure 4 show the amino acid interaction percentages for alanine, glycine, isoleucine, leucine, valine, aspartic acid, serine and threonine in the 5Å and 7Å analysis, respectively. A full table with all the interaction percentages of all amino acids for bacteria can be found in the Appendix.

In reverse, aspartic acid has high interaction percentages with serine and threonine in the 7Å analysis, but with arginine and serine in the 5Å analysis. Medium interaction percentages with asparagine, aspartic acid, as well as threonine, can be observed in both analyses. Serine also has notable interactions with aspartic and glutamic acid, serine and threonine, with similar observations made for threonine in both 7Å and 5Å analysis.

Amino acids like arginine, glutamine, aspartic and glutamic acid also exhibit notable interactions with aspartic acid, serine and threonine, in both analyses. It is of interest that histidine interacts notable with aspartic and glutamic acid, glycine, leucine, serine, threonine and valine, while none of these amino acids maintain notable interactions with histidine. Similar observations can be made for lysine, despite the fact that glutamic acid and lysine show a medium interaction percentage in both analyses. Lastly, cysteine exhibits mostly high interaction percentages with all the hydrophobic residues discussed above (alanine, isoleucine, leucine and valine), but does not exhibit any notable interaction with itself. However, it is of interest that high interaction percentages with arginine and asparagine are observed in 7Å and 5Å analysis, respectively.

3.1.2 Archaea

In archaea, all amino acids showed both high and medium interaction percentages with alanine, glycine, isoleucine, leucine and valine. Specifically, the 7Å analysis showed that all residues interact highly with valine, leucine and isoleucine, when alanine and glycine have both high and medium level interactions, depending on the residue in question. The 5Å analysis showed constantly high interactions for isoleucine and leucine, with alanine and glycine displaying both high and medium interaction percentage patterns.

On the other hand, alanine and valine themselves have a rather small number of notable interactions, when in comparison with their interaction percentage with all amino acids. In the 7Å analysis, alanine and valine interact highly with alanine, isoleucine, leucine and valine, followed by medium interaction percentage with glutamic acid and threonine. In the 5Å analysis, alanine and valine show medium interactions with arginine, lysine and phenylalanine, along with all previous interaction percentages. Isoleucine and leucine share remarkably similar percentages in both analyses, both in quality and quantity. Specifically, both isoleucine and leucine interact highly with alanine, isoleucine, leucine and valine, followed by medium interactions with glutamic acid, glycine, serine and lysine. It is also worth noting that the interaction percentages are almost identical, both when comparing the interactions of the amino acids themselves and when comparing the interaction percentages of all amino acids with isoleucine and leucine. Lastly, glycine displays high interaction with alanine, isoleucine, leucine and valine, followed by medium percentages with arginine, glutamic and aspartic acid, serine and threonine in both analyses.

In terms of polar and charged amino acids, a large number of interaction percentages, both high and medium, are observed for arginine, aspartic and glutamic acid, lysine, serine and threonine, in both analyses. In detail, arginine, asparagine, glutamine, glutamic acid, lysine, serine and threonine show high interaction percentages for aspartic and glutamic acid, followed by medium percentages with alanine, aspartic acid, glycine, methionine, phenylalanine and tryptophan. Aspartic acid, asparagine, proline and tyrosine have high interaction percentages for serine, followed by medium interactions with glutamine, glutamic acid in both analyses. Similar interactions were accounted for threonine. Surprisingly, all amino acids show a notable interaction percentage with arginine in the 5Å analysis, but only a few retain such a notable interaction in the 7Å analysis. Figure 5 and Figure 6 show the amino acid interaction percentages for alanine, glycine, isoleucine, leucine, valine, aspartic and glutamic acid, arginine, serine and threonine in the 5Å and 7Å analysis, respectively. A full table with all the interaction percentages of all amino acids for archaea can be found in the Appendix.

In reverse, arginine, aspartic and glutamic acid have high interaction percentages with arginine, aspartic and glutamic acid in the 5Å analysis, while in the 7Å analysis, arginine and glutamic acid maintain a high percentage of interactions with aspartic acid. Medium interaction percentages with arginine, aspartic and glutamic acid include alanine, glycine and lysine. Furthermore, lysine, serine and threonine have high interaction percentages among and with themselves. Medium percentages with lysine, serine and threonine, can be observed with alanine, glycine and

serine. Lastly, cysteine has notable interactions with alanine, glycine, isoleucine, leucine, valine and cysteine, the latest being the largest of them all.

3.1.3 Eukarya

In eukarya, all amino acids showed both high and medium interaction percentages with alanine, glycine, isoleucine, leucine and valine. Specifically, the 7Å analysis showed that all residues interact highly with leucine, when alanine, glycine, isoleucine and valine displays both high and medium interactions, depending on the residue in question. The 5Å analysis showed constantly high interactions with leucine, with alanine, valine and glycine displaying both high and medium interaction patterns. It is also worth mentioning that the 5Å analysis highlighted phenylalanine as a medium interaction percentage residue for all amino acids.

On the other hand, alanine and valine themselves have a rather small number of notable interactions, when in comparison with their interaction percentage with all amino acids. In the 7Å analysis, alanine and valine interact highly with alanine, isoleucine, leucine and valine, followed by medium interactions with glutamic acid, glycine, proline, serine and threonine. In the 5Å analysis, alanine and valine show medium interactions with arginine, lysine and phenylalanine, along with all previous percentages. Isoleucine and leucine share many interactions, usually both medium or high, but they do not share the same interactivity with all amino acids. While leucine maintains a high interaction percentage from all amino acids, isoleucine is mostly medium interacted by most amino acids. Lastly, glycine interacts highly with alanine, glycine, leucine and valine, followed by medium percentages with arginine, glutamic and aspartic acid, isoleucine, serine and threonine in both analyses.

In terms of polar and charged amino acids, a large number of interaction percentages, both high and medium, are observed for arginine, aspartic and glutamic acid, proline, serine and threonine, in both analyses. In detail, aspartic and glutamic acid constitute a mostly medium interaction preference for most amino acids, in both analyses. Similar conclusions can be drawn for proline, serine and threonine, while arginine is seen as a universal medium interaction preference only in the 5Å analysis. Figure 7 and Figure 8 show the amino acid interaction percentages for alanine, glycine, isoleucine, leucine, valine, aspartic and glutamic acid, arginine, proline, serine and threonine in the 5Å and 7Å analysis, respectively. A full table with all the interaction percentages of all amino acids for eukarya can be found in the Appendix.

In reverse, arginine interacts highly with aspartic acid, glutamic acid, as shown by the 7Å analysis, but presents medium interactions with proline, serine and threonine, which provide similar results for themselves. Aspartic and glutamic acid show high interactions with arginine, lysine, serine and themselves respectively, followed by medium percentages with asparagine, threonine, valine and alanine. It is of interest that histidine, asparagine and glutamine maintain notable interactions with almost all amino acids, while all residues do not maintain notable percentages with the former, in both analyses. Serine and threonine show a wide spectrum of notable interactions, most of them being average and with the aforementioned amino acids. Finally, cysteine interacts highly with alanine, cysteine, isoleucine, leucine and valine, followed by medium percentages with phenylalanine and serine in both analyses.

A unique point of the 7Å analysis is the highlight of proline as a medium interaction preference from most amino acids. It is also remarkable the fact that isoleucine and leucine, which count among the most interactive amino acids, have insignificant interaction percentages with proline.

Figure 3 Logarithmic amino acid interaction percentages for alanine, glycine, isoleucine, leucine, valine, aspartic acid, serine and threonine in bacteria. Analysis performed at 5Å.

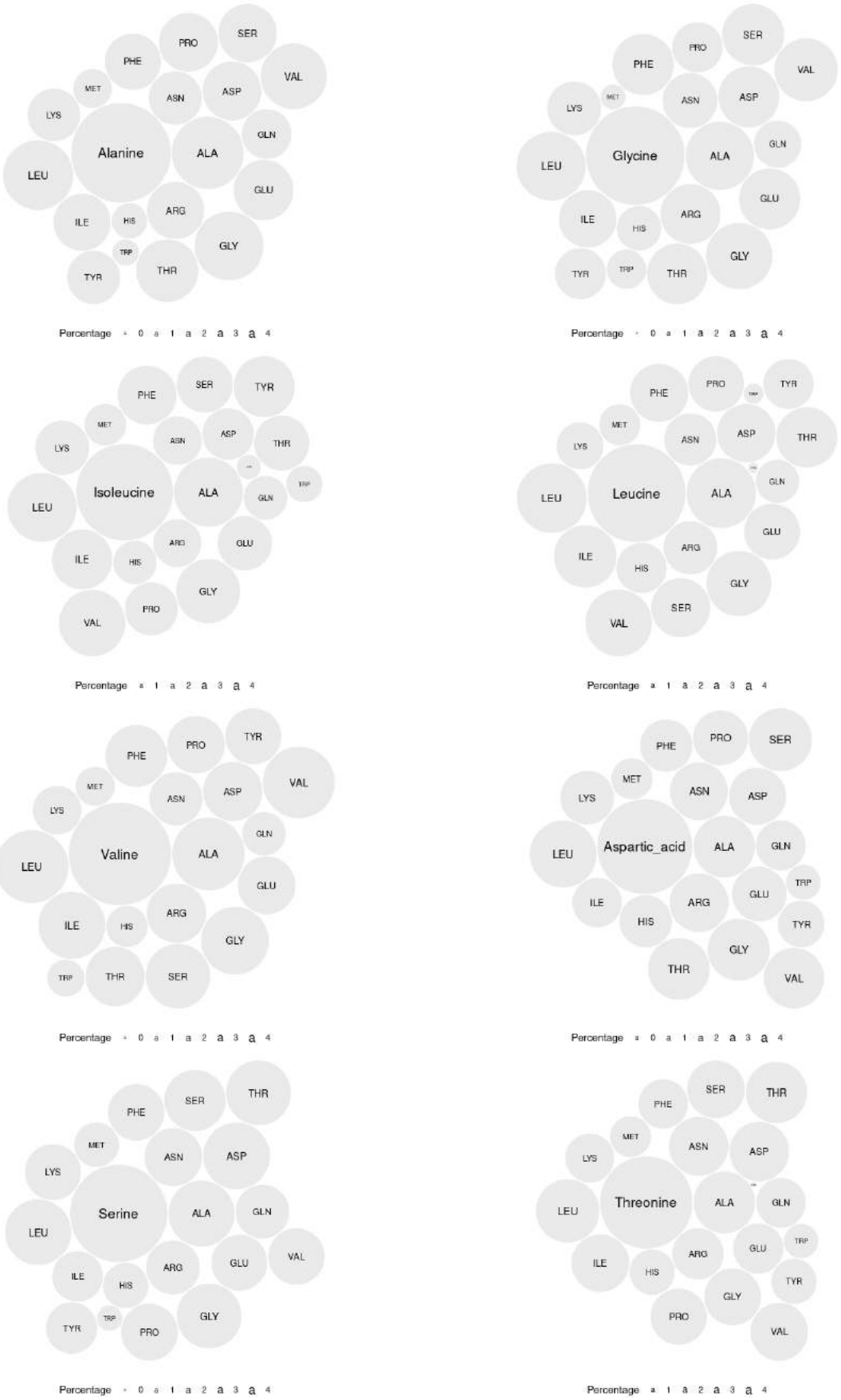
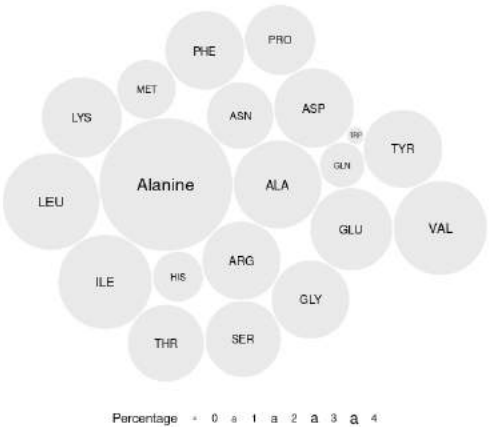
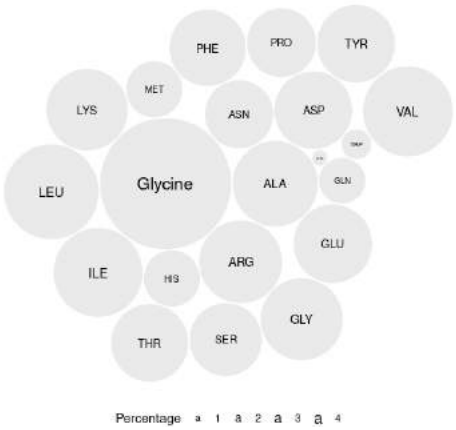
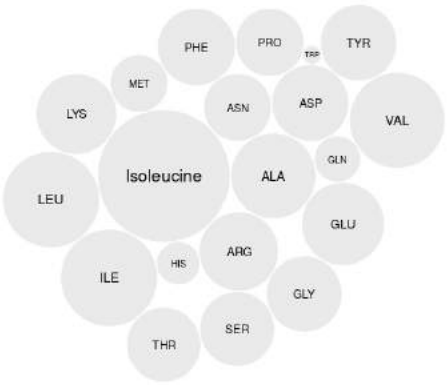
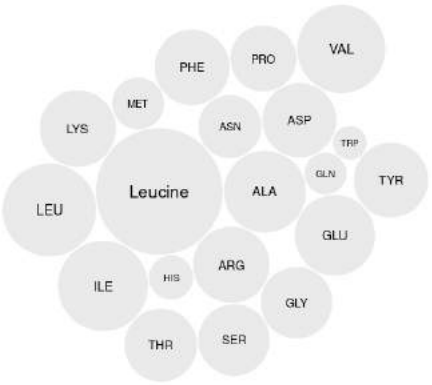


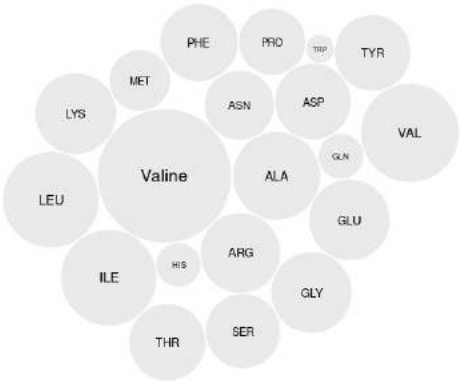
Figure 4 Logarithmic amino acid interaction percentages for alanine, glycine, isoleucine, leucine, valine, aspartic acid, serine and threonine in bacteria. Analysis performed at 7Å.

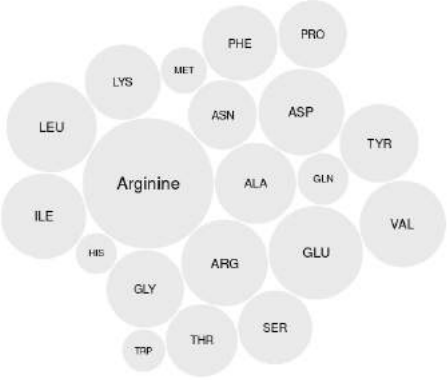


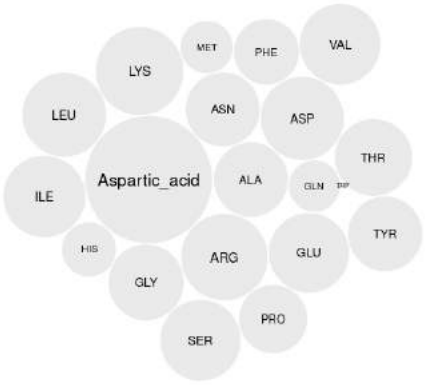












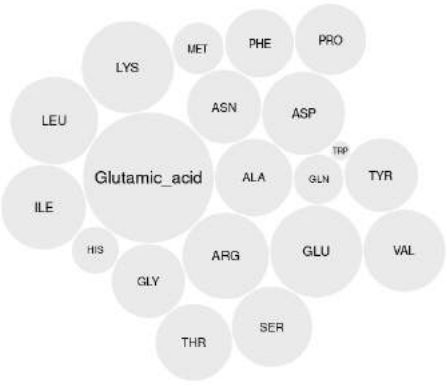
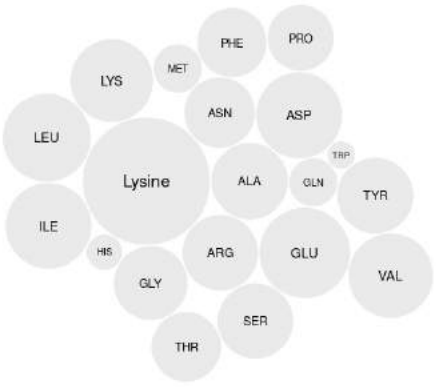
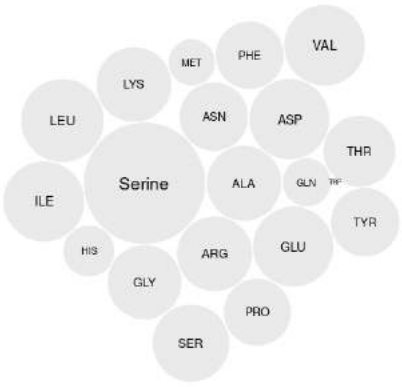


Figure 5 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, lysine, serine and threonine in archaea. Analysis performed at 5Å.(continues in next page)





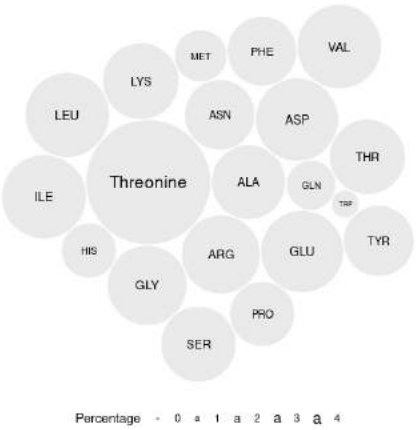


Figure 5 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, lysine, serine and threonine in archaea. Analysis performed at 5Å.

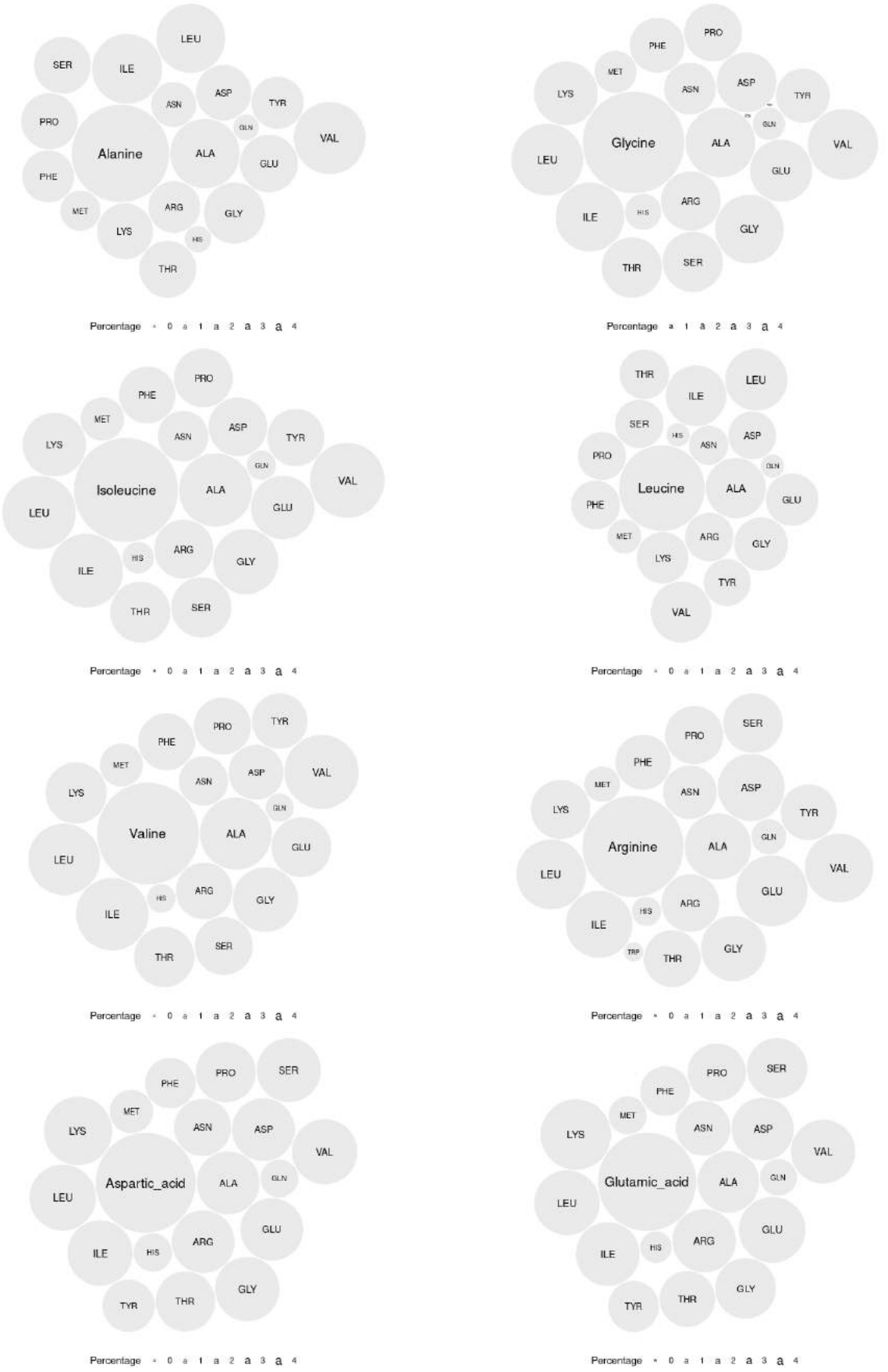


Figure 6 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, lysine, serine and threonine in archaea Analysis performed at 7Å. (continues in next page)

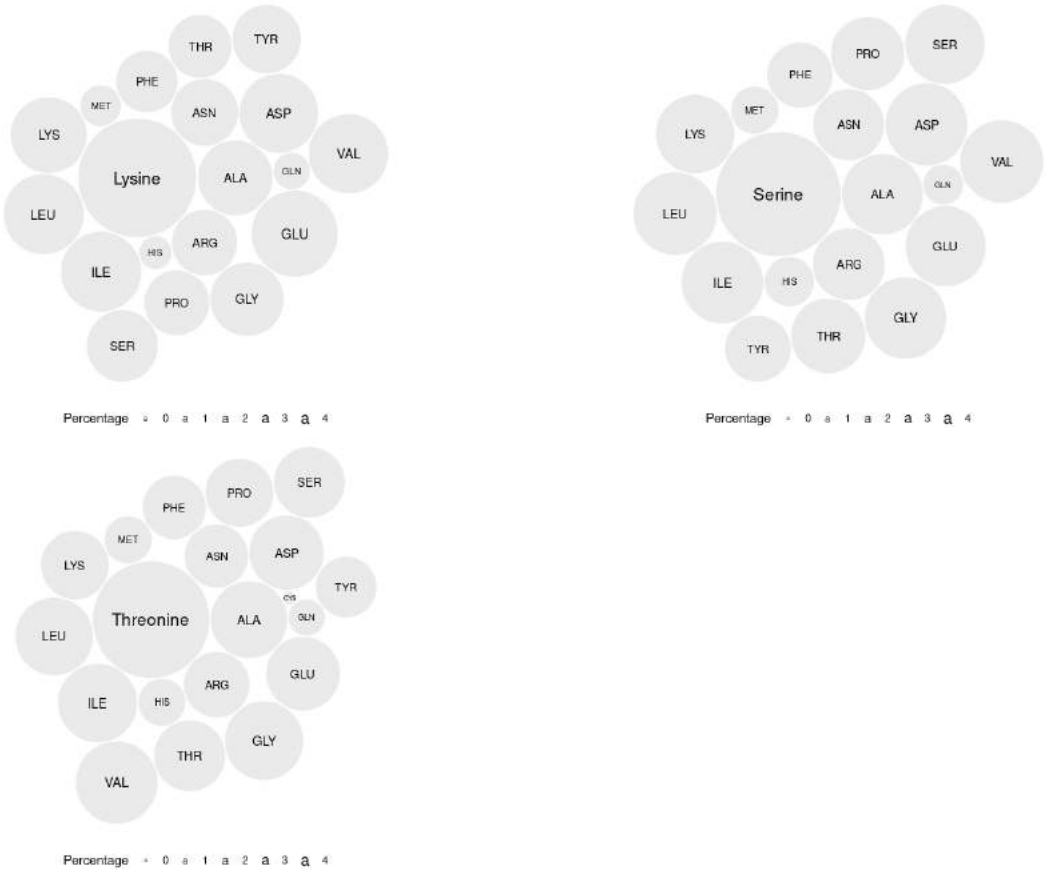
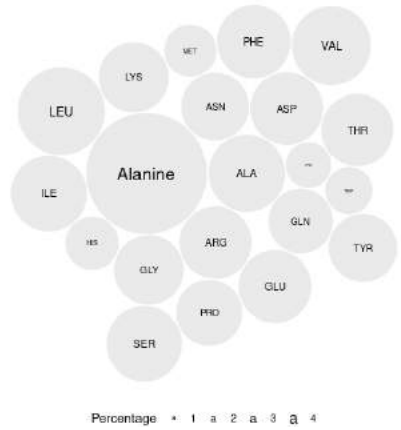
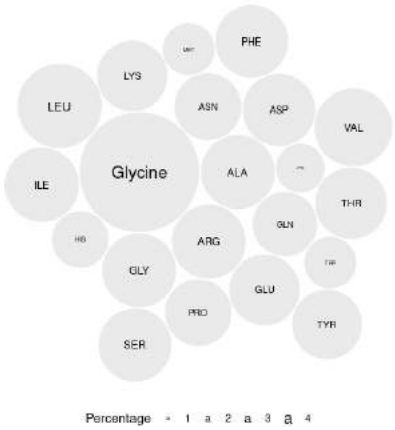
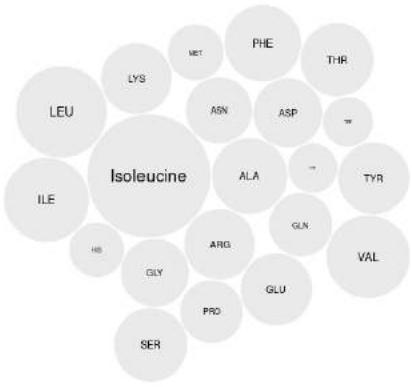
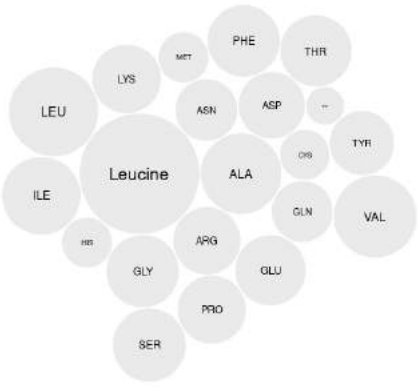


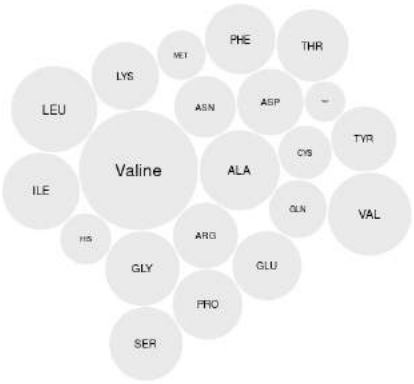
Figure 6 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, lysine, serine and threonine in archaea. Analysis performed at 7Å.

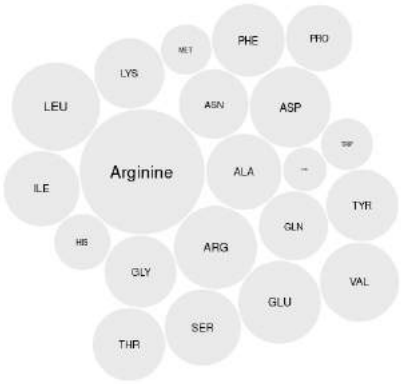


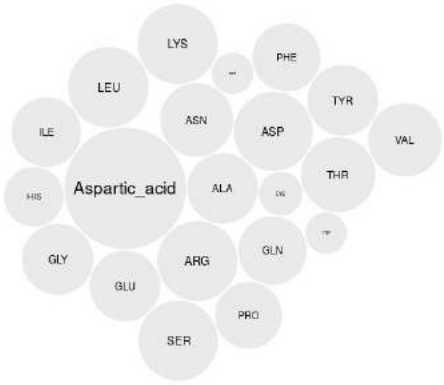












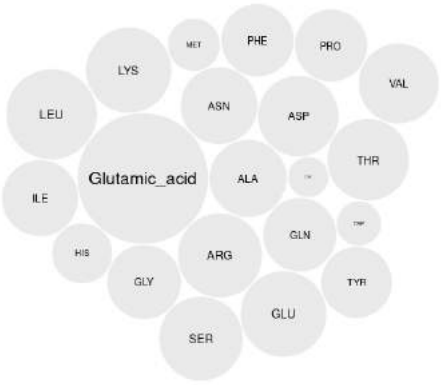


Figure 7 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in eukarya. Analysis performed at 5Å (continues in next page)

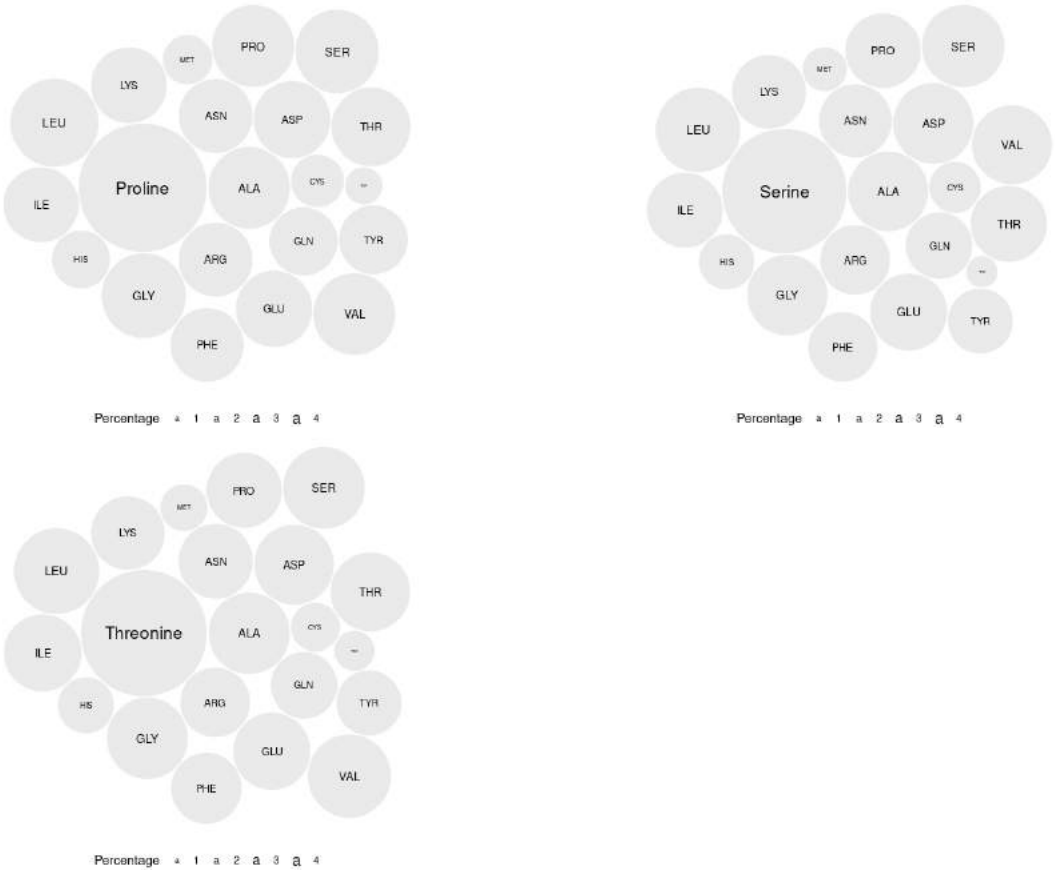
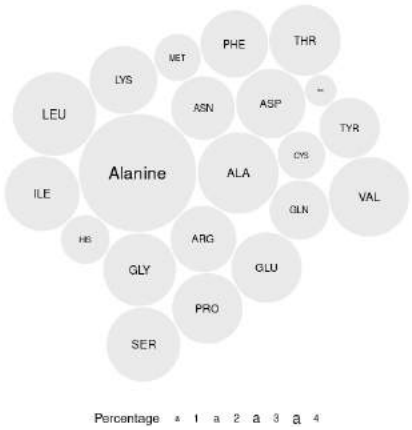
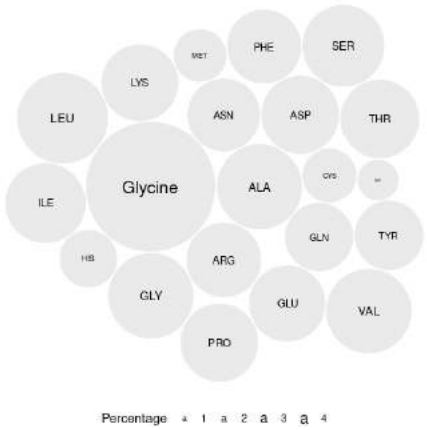
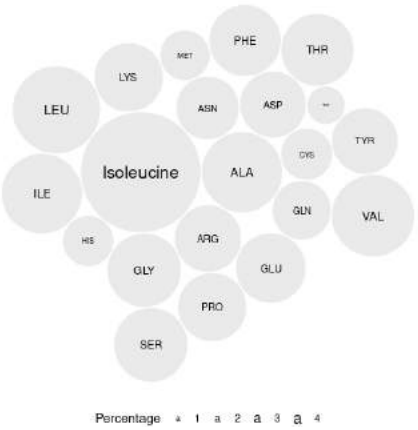
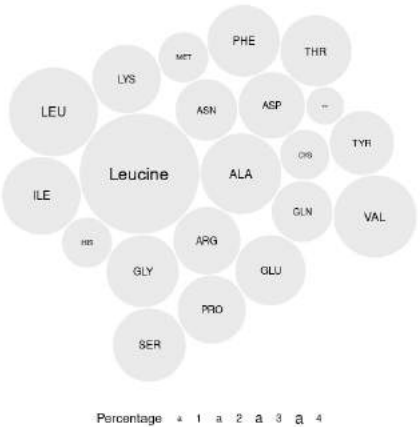


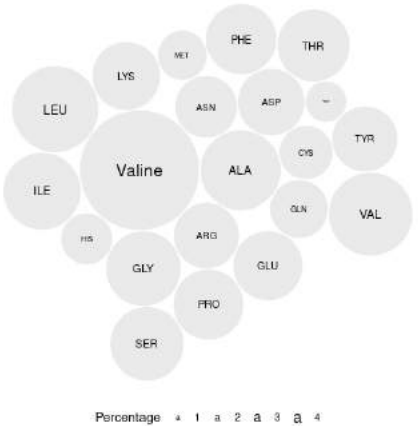
Figure 7 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in eukarya. Analysis performed at 5Å.

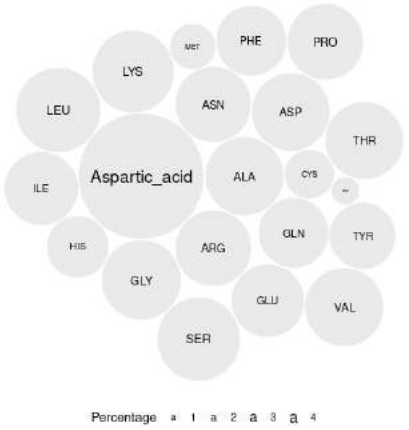












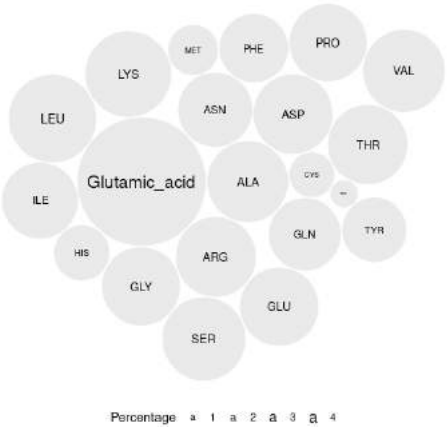


Figure 8 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in eukarya. Analysis performed at 7Å (continues in next page)

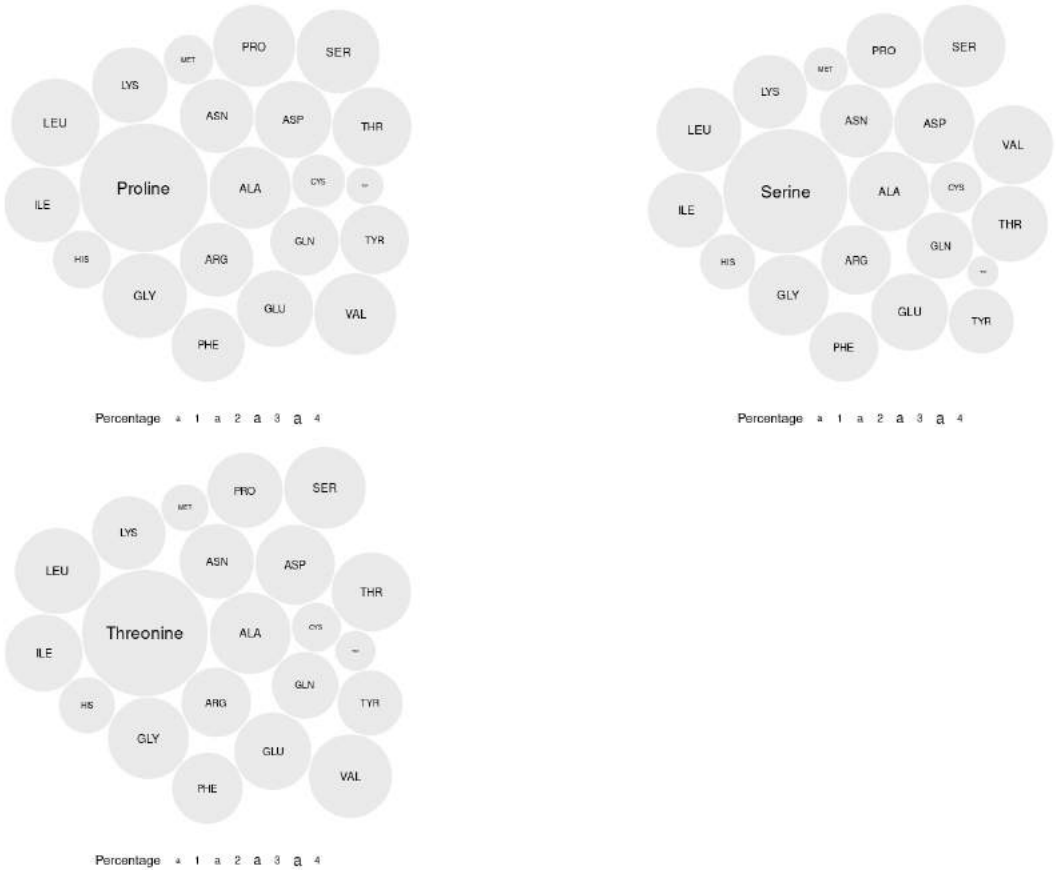


Figure 8 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in eukarya. Analysis performed at 7Å.

3.2 Amino acid interactions in mesophiles and thermophiles

Study of all structures based on temperature has provided interesting results regarding the interaction preferences of the amino acids. The structures were classified by their respective organism's optimal growth temperature in two categories (mesophiles and thermophiles) and then processed by clusters.pl and clusters2_1.pl. Logarithmic interaction percentages were arbitrarily categorised in three classes: high interaction percentages for values greater or equal to 1.86 (26%), medium interaction percentage for values greater or equal to 1.55 (20%) and minute interaction percentages for values lower than 1.55 (20%).

Overall, leucine has been highlighted as a high interaction percentage preference for all amino acids, both in 5Å and 7Å analysis. Isoleucine has been found to be a medium interaction percentage preference, whilst alanine, glycine and valine show a mixed interaction percentage preference pattern, depending on the domain of the organism, as well as the amino acid studied. In terms of polarity and charge, aspartic and glutamic acid seem to be a universally preferred interaction percentage preference for polar and charged amino acids, while cystein has been found to interact mostly with hydrophobic residues and itself. Further interaction percentage preferences are discussed in detail in the following sections. A complete list of all interaction percentage preferences, both in tables and illustrations, is available in the Appendix.

3.2.1 Mesophiles

In mesophiles, all amino acids showed both great and medium interaction percentages with alanine, glycine, isoleucine, leucine, serine, threonine and valine. Specifically, the 7Å analysis showed that all residues interact highly with alanine, leucine and valine, when glycine displays both great and medium interaction percentages, depending on the residue in question. The 5Å analysis showed a constant high interaction percentage pattern for leucine, with

alanine, valine and glycine displaying both high and medium interaction percentage patterns. It is also worth mentioning that the 5Å analysis highlighted arginine, aspartic and glutamic acid as a mostly medium interaction percentage residue for all amino acids.

On the other hand, alanine and valine themselves have a rather small number of notable interactions, when in comparison with their interaction percentage with all amino acids. In the 7Å analysis, alanine and valine interact highly with alanine, glycine, isoleucine, leucine and valine, followed by medium interactions with proline, serine and threonine. Alanine also retains medium interaction levels with aspartic and glutamic acid. In the 5Å analysis, alanine and valine show medium percentages with arginine, glutamic acid, glycine and phenylalanine, along with all previous interactions. Isoleucine and leucine share many interactions, usually both medium or high, but they do not share the same interactivity with all amino acids. While leucine interacts highly with all amino acids, isoleucine is mostly a medium interaction preference for most amino acids. Lastly, glycine displays mostly high interactions with alanine, glycine, leucine and valine, followed by medium percentages with arginine, aspartic and glutamic acid, isoleucine, serine and threonine in both analyses.

In terms of polar and charged amino acids, a large number of interaction percentages, both high and medium, are observed for aspartic and glutamic acid, serine and threonine, in both analyses. In detail, aspartic and glutamic acid constitute a mostly medium interaction preference for most amino acids, in both analyses. Similar conclusions can be drawn for serine and threonine, while arginine is seen as a mostly medium interaction preference only in the 5Å analysis. Figure 9 and Figure 10 show the amino acid interaction percentages for alanine, glycine, isoleucine, leucine, valine, aspartic acid, serine and threonine in the 5Å and 7Å analysis, respectively. A full table with all the interaction percentages of all amino acids for mesophiles can be found in the Appendix.

In reverse, aspartic and glutamic acid show high interactions with arginine, lysine, serine and themselves respectively, followed by medium percentages with asparagine, threonine, valine and alanine. It is of interest that histidine, asparagine and glutamine maintain notable interactions with almost all amino acids, while all amino acids do not maintain notable interactions in return, in both analyses. Serine and threonine show a wide spectrum of notable interaction percentages, most of them being of medium level and with the aforementioned amino acids. Finally, cysteine interacts highly with alanine, cysteine, isoleucine, leucine and valine, followed by medium interaction percentages with phenylalanine and serine in both analyses.

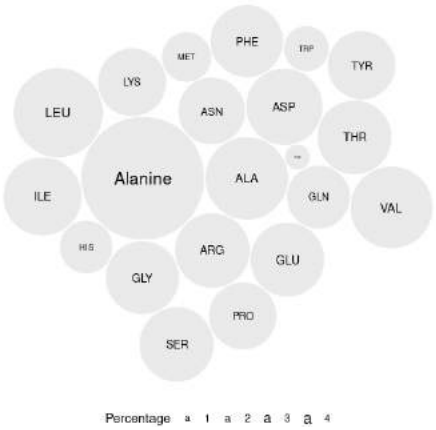
3.2.2 Thermophiles

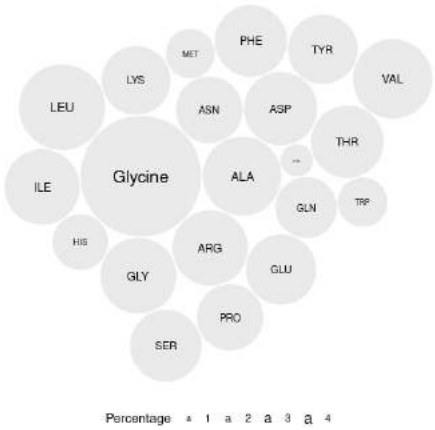
In thermophiles, all amino acids showed both great and medium interaction percentages with alanine, glycine, isoleucine, leucine and valine. Specifically, the 7Å analysis showed that all residues interact highly with leucine and valine, when alanine, glycine and isoleucine display both great and medium interaction percentages, depending on the amino acid in question. The 5Å analysis showed a constant high interaction percentage pattern for valine, with alanine, isoleucine and leucine interacting both highly and in medium levels. It is also worth mentioning that in both analyses, most amino acids interact highly with isoleucine, something that wasn't the case in most previous analyses.

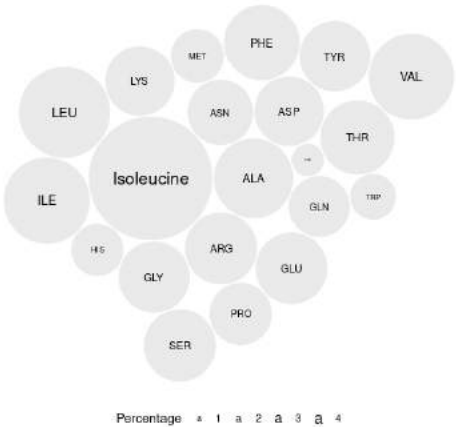
On the other hand, alanine and valine themselves have a rather small number of notable interactions, when in comparison with their interaction percentage with all amino acids. In the 7Å analysis, alanine and valine show high interactions with alanine, glycine, isoleucine, leucine and valine, followed by medium percentages with aspartic and glutamic acid, as well as lysine. In the 5Å analysis, alanine and valine show medium percentages with arginine, aspartic and glutamic acid, glycine, lysine and phenylalanine, along with all previous interactions. Isoleucine and leucine are involved in the formation of the greatest number of contacts, with all residues. However, both analyses highlight leucine as the main high interaction percentage preference for amino acids. Lastly, glycine interacts mostly highly with alanine, glycine, isoleucine, leucine and valine, followed by medium percentages with arginine, aspartic and glutamic acid and threonine in both analyses.

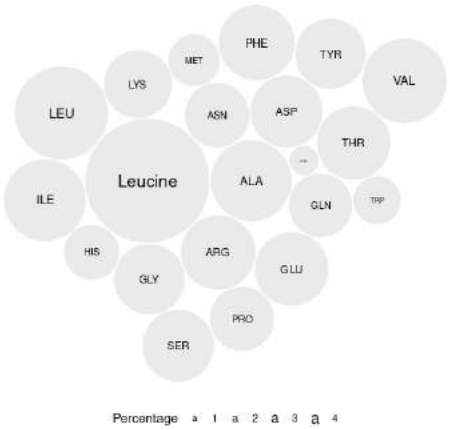
In terms of polar and charged amino acids, a large number of interaction percentages, both high and medium, are observed for arginine, aspartic and glutamic acid, threonine and tyrosine, in the 5Å analysis. Amino acids also show high and medium interaction percentages for glutamic acid and threonine in the 7Å analysis. In detail, arginine, aspartic and glutamic acid constitute a high and medium interaction preference for most amino acids, depending on the residue in question. Threonine is solely a medium interaction preference for most amino acids, in both analyses, while tyrosine expresses the same pattern as threonine only in the 5Å analysis, with a few exceptions.

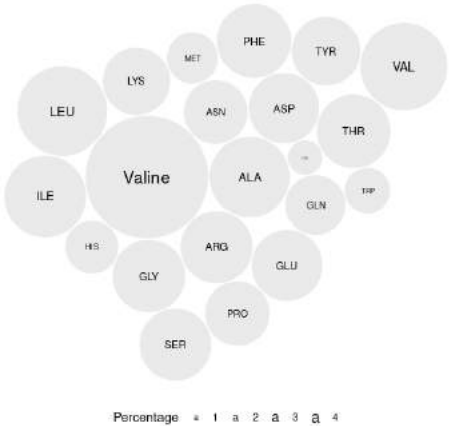
In reverse, aspartic and glutamic acid show high interactions with arginine, lysine, valine and themselves respectively, followed by medium percentages with asparagine, threonine, tyrosine and alanine. It is of interest that histidine, asparagine and glutamine maintain notable interaction percentages with many amino acids, while all residues do not maintain notable percentages with the former, in both analyses. Tyrosine and threonine show a wide spectrum of notable interaction percentages, most of them being average and with the aforementioned amino acids. Finally, cysteine interacts highly with cysteine, glutamic acid, isoleucine, leucine and valine, followed by medium percentages with phenylalanine and tyrosine in both analyses. Figure 11 and Figure 12 show the amino acid interaction percentages for alanine, glycine, isoleucine, leucine, valine, aspartic acid, serine and threonine in the 5Å and 7Å analysis, respectively. A full table with all the interaction percentages of all amino acids for thermophiles can be found in the Appendix.

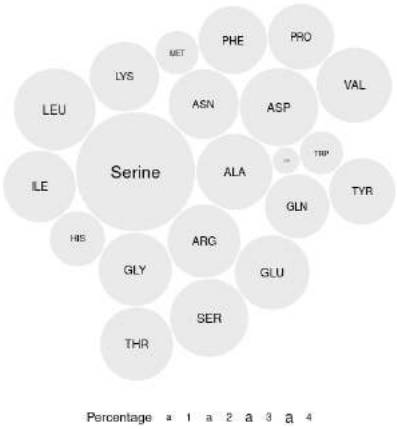


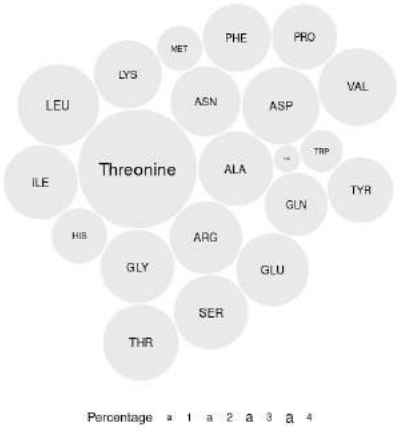












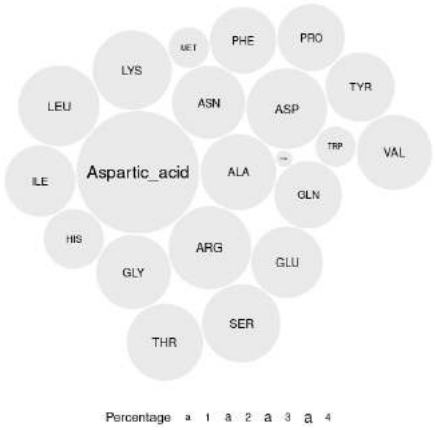


Figure 9 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline serine and threonine in mesophiles. Analysis performed at 5Å (continues in next page)

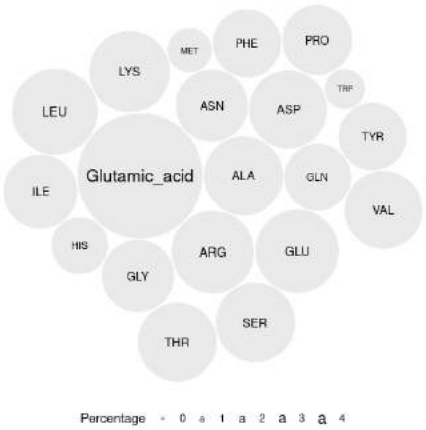


Figure 9 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in mesophiles. Analysis performed at 5Å.

Figure 10 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in mesophiles. Analysis performed at 7Å. (continues in next page)

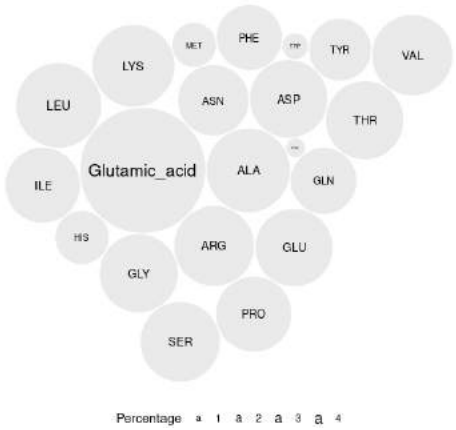
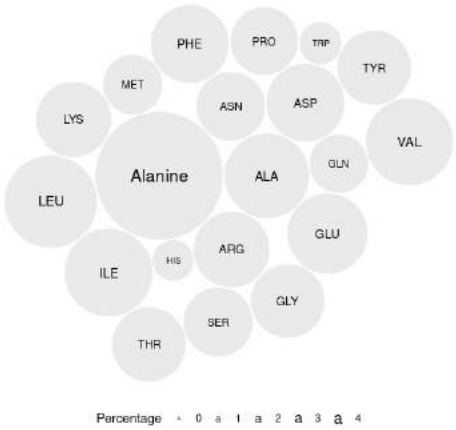
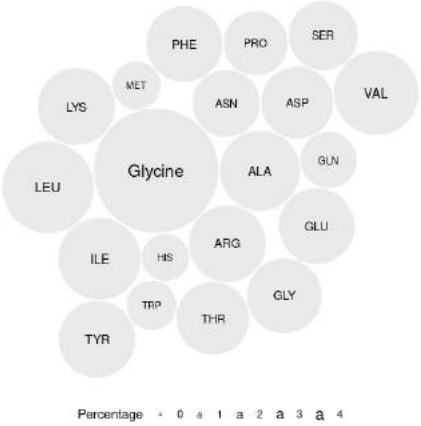
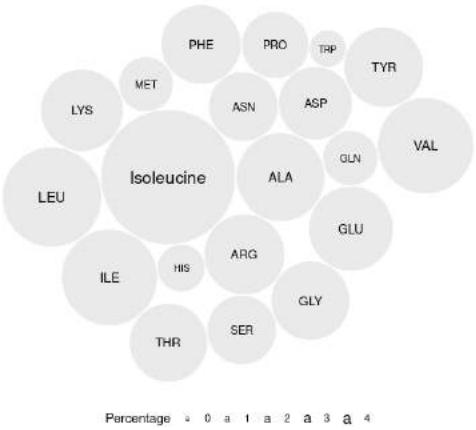
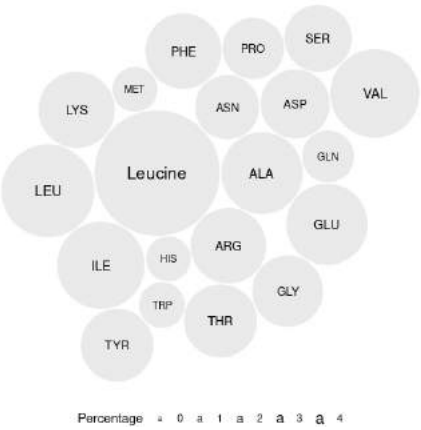


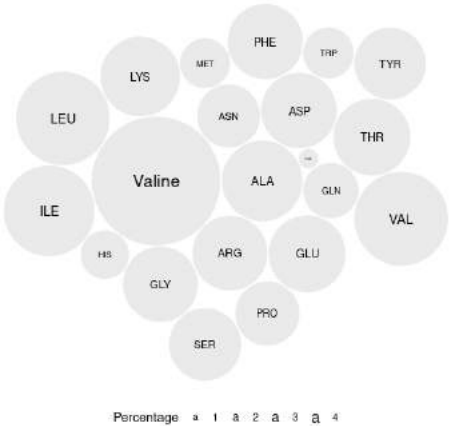
Figure 10 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in mesophiles. Analysis performed at 7Å.

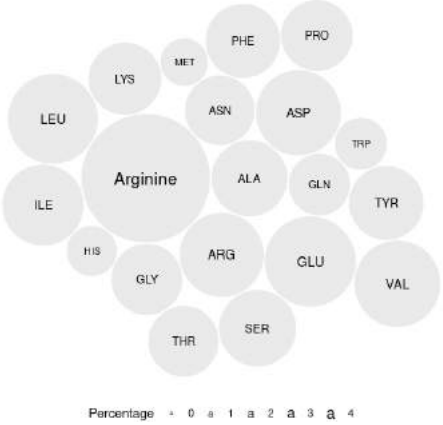


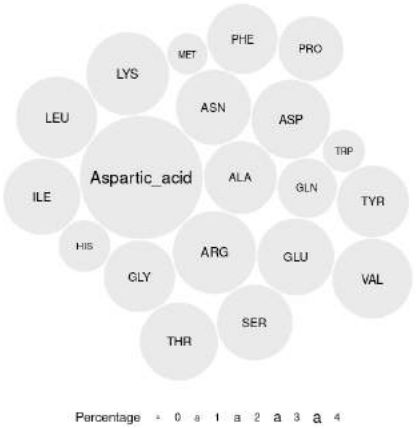












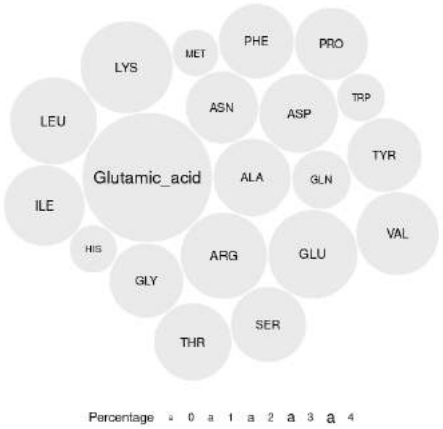


Figure 11 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine, threonine in thermophiles. Analysis performed at 5Å. (cont in next page)

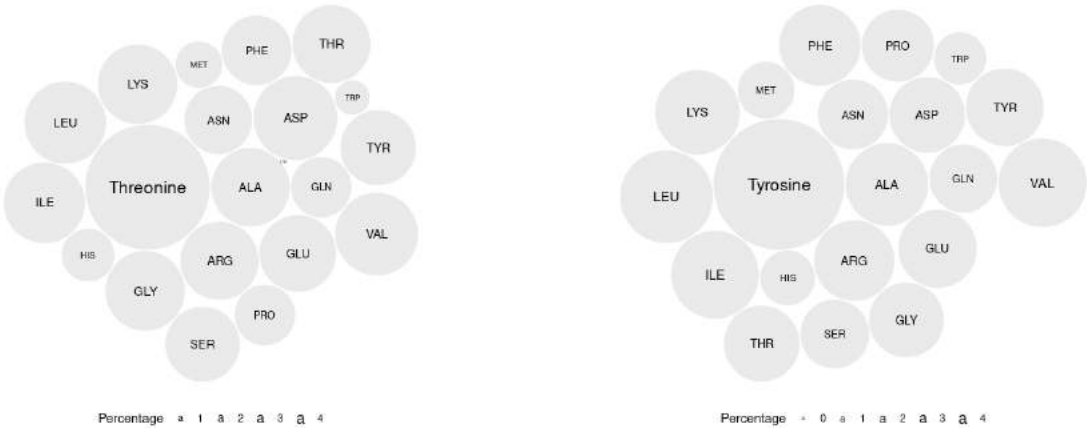
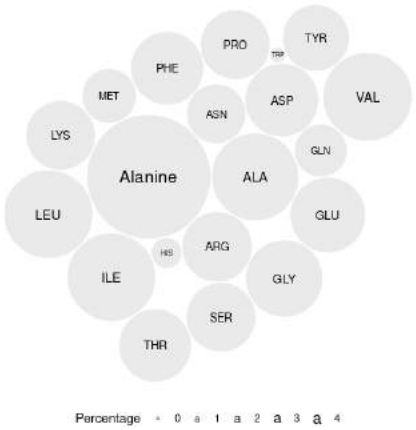
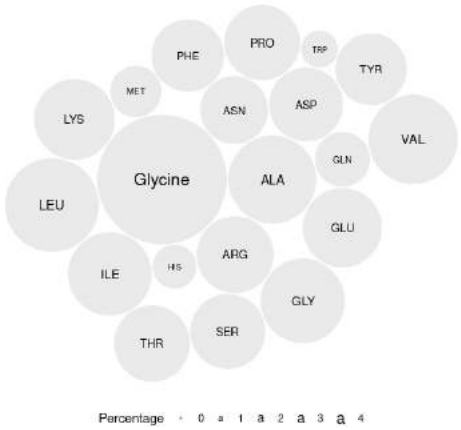
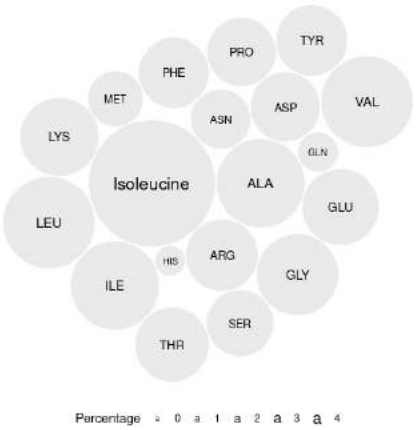
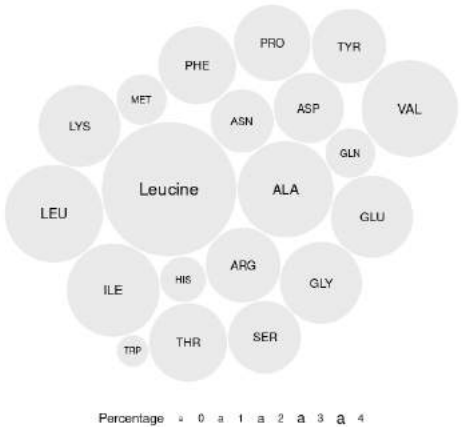


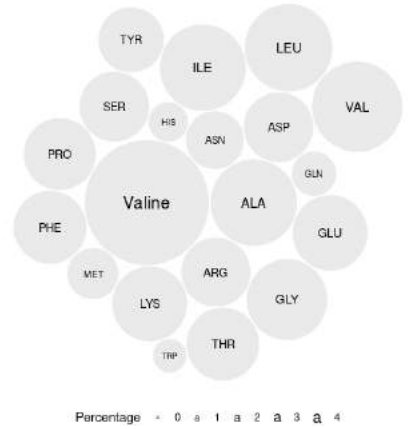
Figure 11 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in thermophiles. Analysis performed at 5Å.

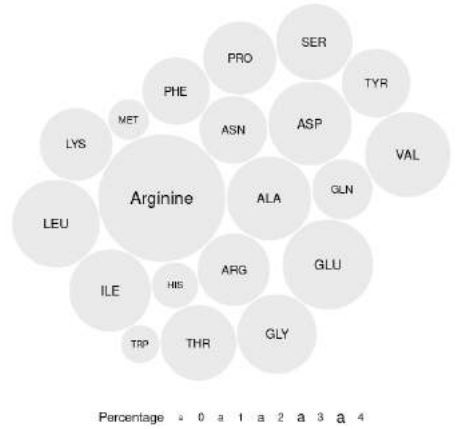


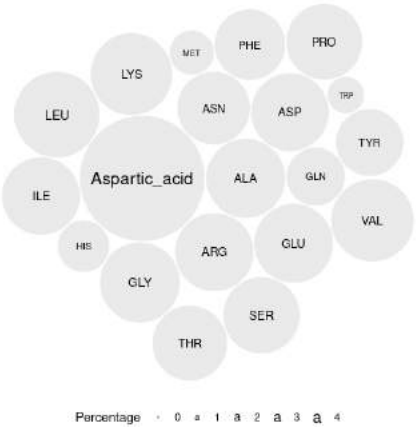












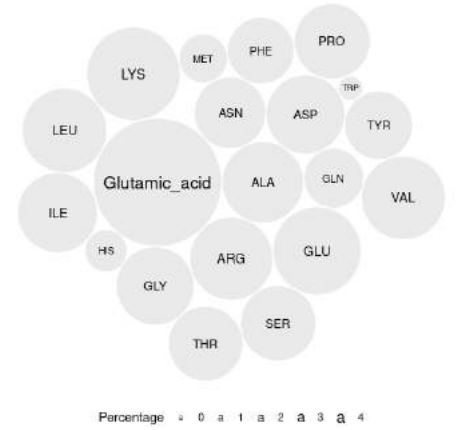


Figure 12 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine, threonine in thermophiles. Analysis performed at 7Å. (cont in next page)

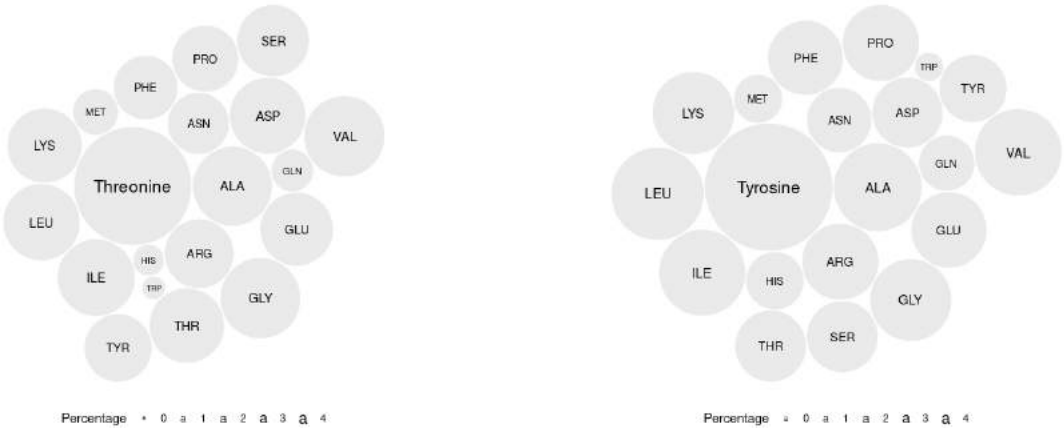


Figure 12 Logarithmic amino acid interaction percentages for alanine, leucine, glycine, isoleucine, valine, arginine, aspartic and glutamic acid, proline, serine and threonine in thermophiles. Analysis performed at 7Å.

These analyses aimed at exploring the possible existence of amino acid interaction patterns and how they are affected by extrinsic effectors. Our analyses suggest that there are conserved patterns of amino acid interactions in globular proteins, which indeed change according to the group of study. First, leucine has been highlighted as the major amino acid for high interaction percentage with all residues, with isoleucine being the second best alternative of interaction choice. Alanine and valine play significant roles as well, with their great and medium interaction percentages. It is also important to note the appearance of aspartic and glutamic acid, as well as arginine and lysine among the most interactive amino acids. These residues are known for being both hydrophobic and hydrophilic at the same time, depending on the interacting segment of their molecule. While cysteine has not been observed to be an interactive residue, its contribution to three-dimensional structure via covalent, disulphide bonds should not be overlooked. Therefore, it was interesting to observe both the high interaction percentages of cysteine with alanine, glycine, isoleucine, leucine and valine, as well as its lack of interaction with itself in the case of bacteria. Another intriguing example is proline, with its appearance as a medium interaction percentage preference for most residues in eukarya.

As such, it could be suggested that alanine and valine play a pivotal role in the formation of the hydrophobic core of the protein, while leucine and isoleucine provide an interaction interface with the hydrophobic core for all residues. Given their branched and long side chain, leucines and isoleucines could be used as a scaffold to connect the hydrophobic core, mainly composed of alanine and valine, with the solvent accessible residues or active centres. Residues with both hydrophobic and hydrophilic segments, like aspartic and glutamic acid could be used to create buried salt bridges and ion pairs, as well as participate in active centres and the catalysis of chemical reactions. The rest of the 20 residues, like histidine, asparagine or glutamine seem to be used and positioned only for very specific purposes, given their large interaction percentages with many of the aforementioned residues, but without receiving any response.

In terms of our studies on temperature, our analyses have highlighted that in thermophiles, there is a much more focused use of amino acids, in comparison with mesophiles. While mesophiles utilise a wider spectrum of the available amino acids in their interactions (16 out 20 amino acids have at least medium interaction percentages for more than 10 residues), thermophiles direct their interactions to specific residues, the most prominent of them being the ones at Figure 11 and Figure 12. A point of question is whether psychrophiles would display similar behaviour, given their extreme temperature.

However, this study is far from refined. Our analysis thus far has suggested the existence of conserved clusters of contacts in the domain level, as well as in response to temperature, but it has little to provide in terms of more specific taxonomic categorisation or other environmental factors. One possible venture would be the differentiation of organisms according to kingdoms or phyla, to see whether our methodology could provide substantial results. Another could be the study of other extrinsic effectors, like salinity, pH or hydrostatic pressure. Furthermore, for our studies, we selected, arbitrarily, the value thresholds for medium and high interaction percentages. Therefore, further analyses on different threshold values could define the optimum values for high and medium interaction percentages. Lastly, these analyses could be used along with other amino acid network or amino acid composition analyses to provide further insight in the structure and formation of proteins.

All these observations could be used to discriminate proteins or, at this stage, proteomes of organisms. For example, an analysis of a proteome with insignificant interaction percentages of cysteine for cysteine could indicate to a bacterial origin, while a medium interaction percentage of all residues with proline could suggest a eukaryote organism. In archaea, we notice an interchanging use of leucine and isoleucine, while this is not the case in the other two domains of life. The direction of the amino acid interactions could be used to distinguish between mesophiles and thermophiles. In the case of benthic sea microorganisms, such methodologies could assist research substantially, given the complex and often overlapping, in regards to extrinsic effectors, natural habitats of these forms of

life. Therefore, such analyses could prove useful as a supplement for major analyses in the classification and study of unknown organisms, for whom very little information is known of.

5. conclusions

Protein structure is the summary of amino acid interactions and environmental effectors. Despite serving numerous functions or adopting various structures and shapes, proteins acquire their functionality through interactions among their residues and with their environment, organised at the various levels of protein structure. For this purpose, we developed a methodology for analysing proteins at their amino acid interaction level, using the network theory approach and the respective amino acid networks theory.

To this end, we constructed 2 analyses, one focusing on the backbone of the protein and another focusing on all interactions among atoms of a protein. The backbone analysis is a 7Å unweighted amino acid network, while its all-atom counterpart is a 5Å weighted amino acid network, prioritising hydrophobic interactions and Coulomb forces. Protein structures were downloaded from the Protein Data Bank, using PISCES and our analysis scripts. Studies were performed in regards to the taxonomy classification of proteins, as well as the temperature spectrum, focusing on mesophiles and thermophiles, due to insufficient structures from psychrophiles.

Our analyses showed a set of conserve d amino acid interactions in globular proteins, along with various shifts in preferences, according to the group of study. In all proteins, leucine and isoleucine interact highly with all residues, while alanine and valine exhibit both great and medium interaction percentages with most amino acids. Various polar and charged residues displayed a continuous appearance in medium interaction percentages, like aspartic and glutamic acid, while residues like histidine, asparagine and glutamine were highly interactive with most residues, but not receiving the same level of interaction. Lastly, cysteine had interestingly low interaction percentages with itself in bacteria, while proline was highlighted as a potential distinguishing factor for proteins or proteomes of eukaryotic origin, thanks to its medium interaction percentage with all residues.

These analyses though, are only a primary study on the field of amino acid interactions. There is a great number of possible analyses to be performed, which will be able to provide further insight into the formation and structure of proteins. Namely, we would like to mention further investigations in multiple environmental factors, like pH, salinity and hydrostatic pressure, as well as performance of our analyses in parallel with amino acid composition methodologies in both evolutionary studies and the study of unknown organisms. Lastly, these analyses could be utilised to provide a distinction and classification methodology for proteomes.

references

Pro Perl: A comprehensive guide for developers who want to master the Perl progamming language
Wainwright P, *Apress*, 2005, ISBN (pbk) 1-59059-438-X

Sams Teach Yourself Perl in 24 Hours, 3rd edition
Pierce C, *Sams Publishing*, Indianapolis, IN, USA
ISBN 0-672-32793-7

The construction of an amino acid network for understanding protein structure and function
Yan W et al., *Amino Acids*, 2014, 46(6):1419-1439

PISCES: a protein sequence culling server
Wang G, Dunbrack R, *Bioinformatics*, 2003, 19(12):1589-1591

The Protein Data Bank
Berman H et al., *Nucleic Acids Research*, 2000, 28(1):235-242

Worldwide Protein Data Bank (wwPDB): A virtual treasure for research in biotechnology
Bahzadi P and Gajdács M, *European Journal of Microbiology and Immunology*, 2021, 11(4):77-86

Ανάλυση δεδομένων με την R
Νικολάου Χ, Δίσιγμα, Θεσσαλονίκη, 2019

Introduction to protein structure, 2nd edition
Branden C, Tooze J., *Garland Publishing Inc.*, New York, NY, USA, ISBN (pbk) 0-8153-2305-0

Biochemistry, 8th edition
Berg JM, Tymozcko JL, Gatto GJ, Stryer L., *W.H. Freeman and Co.*, 2015, ISBN 978-960-524-495-8

Principles of Proteomics, 2nd edition
Twyman RM, *Garland Science, Taylor and Francis Group*, 2014, ISBN 978-0-8153-4472-8

Biochemistry, 6th edition
Garrett RH, Grisham CM, *Cengage Learning*, 2017, USA
ISBN 978-1-305-57720-6

Lehninger, Principles of Biochemistry, 7th edition
Nelson DL, Cox MM, *W.H. Freeman and Co.*, 2017
ISBN 978-1-4641-2611-6

Uncovering protein structure
Stollar EJ and Smith DP., *Essays in Biochemistry*, 2020, 64(4):649-680

The origin, evolution and structure of the protein world
Caetano-Anollés G et al., *Biochemical Journal*, 2009, 417:621-637

Hydrogen Bonds: Simple after All?
Herschlag D and Pinney MM., *Biochemistry*, 2018, 57:3338-3352

Low barrier hydrogen bonds in protein structure and function
Trent Kemp M, Lewandowski EM, Chen Y., *BBA – Proteins and Proteomics*, 2021, 140557

Hydrophobic residues can identify native protein structures
Mirzaie M., *Proteins*, 2018, 86:467-474

Water structure and interactions with protein surfaces
Raschke TM., *Current opinion in Structural biology*, 2006, 16:152-159

Salt-bridges in the microenvironment of stable protein structures
Bandyopadhyay AK, Ul Islam RN, Hazra N., *Bioinformation*, 2020, 16(11):900-909

Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings
Bosshard HR, Marti DN, Jelesarov I., *Journal of Molecular Recognition*, 2004, 17:1-16

Driving Protein Conformational Cycles in Physiology and Disease: "Frustrated" Amino Acid Interaction Networks Define Dynamic Energy Landscapes
D'Amico RN, Murray AM, Boehr DD., *BioEssays*, 2020, 42:2000092

Towards an integrated understanding of the structural characteristics of protein residue networks
Khor S., *Theory in Biosciences*, 2012, 131:61-75

The adaptive nature of protein residue networks
Karain WI and Qaraeen NI., *Proteins*, 2017, 87:917-923

Protein structure networks
Greene LH., *Briefings in Functional Genomics*, 2012, 2(6):469-478

Creative elements: network-based predictions of active centers in proteins and cellular and social networks
Csermely P., *Trends in Biochemical Sciences*, 2008, 33(12):569-576

Residue interaction network analysis of Dronpa and a DNA clamp
Hu G et al., *Journal of Theoretical Biology*, 2014, 348:55-64

Protein Contact Networks: An emerging paradigm in chemistry
Di Paola et al., *Chemical Reviews*, 2013, 113:1598-1613

Evolution of protein structures and interactions from the perspective of residue contact networks
Zhang X, Perica T, Teichmann SA., *Current Opinion in Structural Biology*, 2013, 23:954-963

Amino acid interaction networks provide a new lens for therapeutic antibody discovery anr anti-viral drug optimization
Viswanathan K et al., *Current Opinion in Virology*, 2015, 11:122-129

Extremophiles Handbook
Horikoshi K et al., *Springer*, Volume 1, 2011
ISBN 978-4-431-53897-4

The Molecular Basis for Life in Extreme Environments
Ando N et al., *Annual Reviews of Biophysics*, 2021l, 50:343-372

Protein stability and molecular adaptation to extreme conditions
Jaenicke R., *European Journal of Biochemistry*, 1991, 202:715-728

How do thermophilic proteins deal with heat?
Kumar S and Nussinov R., *Cellular and Molecular Life Sciences*, 2001, 58:1216-1233

Thermophilic Adaptation of Proteins
Sterner R and Liebl W., *Critical Reviews in Biochemistry and Molecular Biology*, 2001, 36(1):39-106

Thermophiic Proteins as Versatile Scaffolds for Protein Engineering
Finch AJ and Kim JR, *Microorganisms*, 2018, 6(4):97-110

Lessons in stability from thermophilic proteins
Razvi A and Scholtz JM, *Protein Science*, 2006, 15:1569-1578

Thermophilic proteins: insight and perspective from in silico experiments
Sterpone F and Melchionna S, *Chem. Soc. Rev.*, 2012, 41:1665-1676

Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey
Szilágyi A and Závodszky P, *Structure*, 2000, 8(5):493-504

PERL and R scripts

download.pl

```
use warnings;
use Net::FTP;

#This script (download.pl) downloads all the necessary packages, as provided by
#PISCES. It then calls the contacts.pl script to calculate all the contacts between
#atoms.

#---------COMMANDS---------#

#Connect to ftp.wwpdb.org and log in. Then, go to the right directory and start
#downloading the required packages.
$ftp=Net::FTP->new("ftp.wwpdb.org",Passive=>1,Debug=>0)||die "Unable to connect.\n";
print "Connection established: ftp.wwpdb.org.\n";
$ftp->login("anonymous","\n")||die"Unable to log in.\n"; print "Log in successful.\n";
$fdirectory="/pub/pdb/data/structures/all/pdb/";
$ftp->cwd($fdirectory);
print "Process initiated...\n"; #Let users know that everything is going as expected.

#Firstly, you need to have inserted a PISCES list next to the execution command. Make
#sure to keep only the filename. Then open the provided list.
@ARGV==1||die"Please enter a list next to the execution command.\n";
$ARGV[0]=~/(cull.*)/;
open(LIST, "$ARGV[0]")||die"Unable to read file\n";

#For every line read, find the name of the package and then download it. Then, call
#contacts.pl to calculate the contacts of each amino acid in this package.
while($input=<LIST>)
{
    if($input=~/^(\w\w\w\w)/ && $1!~/PDBc/)
    {
        $code=lc($1); print "\tDownloading pdb$code.ent.gz ... "; #Let users know
        what is downloaded.
        $ftp->get("pdb$code.ent.gz")||warn"File $1 could not be downloaded."; print
        "\n";
        system("gunzip --force pdb$code.ent.gz; perl contacts.pl pdb$code.ent");}}
close(LIST);

#Once downloading has ended, exit ftp.
$ftp->quit;
```

contacts.pl

```
use warnings;

#This script (contacts.pl) calculates all distances between atoms in a .pdb file. At
#first, it extracts data from the .pdb files, then exports contacts lower or equal to
#5 Angstrom. The file exported is a total_contacts_<4-letter-pdb-name>.txt.

#---------COMMANDS---------#

#First, enter the necessary packages for contacts.pl to analyse. Bear in mind that,
#since these packages come from a PISCES list, the file extension is .ent, not .pdb.
(@ARGV==1)||die"Wrong format. Please enter a .pdb file next to the execution command.\
n";
open(PDBFILE, $ARGV[0])||die"There was an error accessing this file.\n";
$ARGV[0]=~/pdb(\w+)\.ent/;
$name=$1;

#We don't need the entire file. We just need the atom coordinates. So, we are going to
#search the file for all lines containing "ATOM" and its respective coordinates. These
#lines will be saved in an array.
@array=();
$index=0;
while($line=<PDBFILE>)
{
    if($line =~ /^ATOM/)
    {
      $array[$index]=$line;
      $index++;}}
close(PDBFILE);

#It is possible that a non-.pdb file is provided. In this case, all procedures should
#be terminated.
if($index==0)
{
    print "$ARGV[0] does not contain a .pdb file.\n";
    exit(0);}

#Assuming that our file is indeed a .pdb file (or that it at least contains the
#necessary data for our analysis), we begin processing. At first, we are going to
#create the export file, named total_contacts_<4-letter-pdb-name>.txt.
open(FINALOUT, ">>total_contacts_$name.txt")||die"There was an error exporting the
data.\n";

#We are going to search the array containing all of the "ATOM" lines. We are going to
#load the data from the first "ATOM" line and then we are going to temporarily load
#the data from the other "ATOM" lines, so we can calculate their distance. Once this
#process has ended, we are going to move to the next "ATOM" line and start over.
$max=$index;
$nocontacts=0;
print FINALOUT "[PDB FILE] $ARGV[0]\n\n"; #Print the name of the pdb file, as it will
be of use in later stages of analysis.
```

```
for($index=0;$index<$max;$index++)
{
    #Save the coordinates of the first amino acid
    if($array[$index]=~/^ATOM\s*\d*\s*(\w+)\s*(\w+)\s*(\w+)\s*(\d+)\s*(-?\d+\Q.\E\d\d\
    d)\s*(-?\d+\Q.\E\d\d\d)\s*(-?\d+\Q.\E\d\d\d)/)
    {
        $identity=$1;
        $amino=$2;
        $chain=$3;
        $no=$4;
        $x=$5;
        $y=$6;
        $z=$7;
        #Start researching the array (with another variable)
        for($jndex=0;$jndex<$max;$jndex++)
        {
            #Save the coordinates of the second amino acid
            if($array[$jndex]=~/^ATOM\s*\d*\s*(\w+)\s*(\w+)\s*(\w+)\s*(\d+)\s*(-?\d+\
                Q.\E\d\d\d)\s*(-?\d+\Q.\E\d\d\d)\s*(-?\d+\Q.\E\d\d\d)/)
            {
                #Calculate their distance
                $distance=sqrt(($5-$x)*($5-$x)+($6-$y)*($6-$y)+($7-$z)*($7-$z));
                #Checking parametres: atoms should not be either in the same amino
                #acid, or neighbouring ones (the amino acids to the left and right of
                #the first amino acid checked). Also, distance should be less or equal
                #to 5 Angstroms. For CA analysis, distance is less or equal to 7
                #Angstrom.
                if($distance<=5 && (($no<$4-1||$no>$4+1) || $chain ne $3))
                {
                    print FINALOUT "$identity $amino $chain $no - $1 $2 $3 $4 //
                 $distance Angstrom\n";
                    $nocontacts++;}}}

        if($nocontacts==0)
        {
            print FINALOUT "$identity $amino $chain $no - No contacts\n";}
        $nocontacts=0;
        print FINALOUT "\n";}}
close(FINALOUT);
print "$ARGV[0] done\n";
exit(1);
```

pcontacts.pl

```
use warnings;

#This script (pcontacts.pl) filters the total_contacts_<4-letter-pdb-name>.txt file,
#produced by contacts.pl, based on physical chemical properties of amino acids.
#Specifically, it filters contacts, based on electrostatic charge and hydrophobic
#interactions. The filtered lines are exported to a ptotal_contacts_<4-letter-pdb-
#name>.txt.

#---------HASH ARRAYS---------#

#This array (%aminos) contains all the atoms in every amino acid to be taken into
#consideration, when checking for preferred contacts. Every amino acid has its own
#hydrophobic atoms, labeled as H (e.g. GLYH, glycine hydrophobic) and polar atoms,
#labeled as P (e.g. GLYP, glycine polar).
%aminos=("GLYH"=>"CA ","GLYP"=>"N O ","ALAH"=>"CB ","ALAP"=>"N O ","VALH"=>"CB CG1 CG2
","VALP"=>"N O ","LEUH"=>"CB CG CD1 CD2 ","LEUP"=>"N O ","ILEH"=>"CB CG1 CG2 CD1
","ILEP"=>"N O ","METH"=>"CB CG CE ","METP"=>"N O ","PHEH"=>"CB CG CD1 CD2 CE1 CE2 CZ
","PHEP"=>"N O ","TRPH"=>"CB CG CD1 CD2 CE2 CE3 CZ2 CZ3 CH2 ","TRPP"=>"N O NE1
","PROH"=>"CB CG CD ","PROP"=>"N O ","SERH"=>"CB ","SERP"=>"N O OG ","THRH"=>"CB CG2
","THRP"=>"N O OG1 ","CYSH"=>"CB ","CYSP"=>"N O SG ","TYRH"=>"CB CG CD1 CD2 CE1 CE2 CZ
","TYRP"=>"N O OH ","ASNH"=>"CB CG ","ASNP"=>"N O OD1 ND2 ","GLNH"=>"CB CG CD
","GLNP"=>"N O OG1 NE2 ","ASPH"=>"CB CG ","ASPP"=>"N O OD1 OD2 ","GLUH"=>"CB CG CD
","GLUP"=>"N O OG1 OE2 ","LYSH"=>"CB CG CD CE ","LYSP"=>"N O NZ ","ARGH"=>"CB CG CD
","ARGP"=>"N O NE NH1 NH2 ","HISH"=>"CB CG CD2 CE1 ","HISP"=>"N O ND1 NE2 ");

#This array (%pairs) contains all of the accepted amino acid interactions. Both ways
#of describing the amino acid pair (e.g. ARGLYS and LYSARG) are included.
%pairs=("GLYALA"=>1,"ALAGLY"=>1,"GLYVAL"=>1,"VALGLY"=>1,"GLYLEU"=>1,"LEUGLY"=>1,"GLYIL
E"=>1,"ILEGLY"=>1,"GLYMET"=>1,"METGLY"=>1,"GLYPHE"=>1,"PHEGLY"=>1,"GLYTRP"=>1,"TRPGLY"
=>1,"GLYPRO"=>1,"PROGLY"=>1,"GLYSER"=>1,"SERGLY"=>1,"GLYTHR"=>1,"THRGLY"=>1,"GLYCYS"=>
1,"CYSGLY"=>1,"GLYTYR"=>1,"TYRGLY"=>1,"GLYLYS"=>1,"LYSGLY"=>1,"GLYARG"=>1,"ARGGLY"=>1,
"GLYHIS"=>1,"HISGLY"=>1,"GLYGLY"=>1,"ALAVAL"=>1,"VALALA"=>1,"ALAALA"=>1,"ALALEU"=>1,"L
EUALA"=>1,"ALAILE"=>1,"ILEALA"=>1,"ALAMET"=>1,"METALA"=>1,"ALAPHE"=>1,"PHEALA"=>1,"ALA
TRP"=>1,"TRPALA"=>1,"ALAPRO"=>1,"PROALA"=>1,"ALASER"=>1,"SERALA"=>1,"ALATHR"=>1,"THRAL
A"=>1,"ALACYS"=>1,"CYSALA"=>1,"ALATYR"=>1,"TYRALA"=>1,"ALALYS"=>1,"LYSALA"=>1,"ALAARG"
=>1,"ARGALA"=>1,"ALAHIS"=>1,"HISALA"=>1,"VALVAL"=>1,"VALLEU"=>1,"LEUVAL"=>1,"VALILE"=>
1,"ILEVAL"=>1,"VALMET"=>1,"METVAL"=>1,"VALPHE"=>1,"PHEVAL"=>1,"VALTRP"=>1,"TRPVAL"=>1,
"VALPRO"=>1,"PROVAL"=>1,"VALSER"=>1,"SERVAL"=>1,"VALTHR"=>1,"THRVAL"=>1,"VALCYS"=>1,"C
YSVAL"=>1,"VALTYR"=>1,"TYRVAL"=>1,"VALLYS"=>1,"LYSVAL"=>1,"VALARG"=>1,"ARGVAL"=>1,"VAL
HIS"=>1,"HISVAL"=>1,"LEULEU"=>1,"LEUILE"=>1,"ILELEU"=>1,"LEUMET"=>1,"METLEU"=>1,"LEUPH
E"=>1,"PHELEU"=>1,"LEUTRP"=>1,"TRPLEU"=>1,"LEUPRO"=>1,"PROLEU"=>1,"LEUSER"=>1,"SERLEU"
=>1,"LEUTHR"=>1,"THRLEU"=>1,"LEUCYS"=>1,"CYSLEU"=>1,"LEUTYR"=>1,"TYRLEU"=>1,"LEULYS"=>
1,"LYSLEU"=>1,"LEUARG"=>1,"ARGLEU"=>1,"LEUHIS"=>1,"HISLEU"=>1,"ILEILE"=>1,"ILEMET"=>1,
"METILE"=>1,"ILEPHE"=>1,"PHEILE"=>1,"ILETRP"=>1,"TRPILE"=>1,"ILEPRO"=>1,"PROILE"=>1,"I
LESER"=>1,"SERILE"=>1,"ILETHR"=>1,"THRILE"=>1,"ILECYS"=>1,"CYSILE"=>1,"ILETYR"=>1,"ILE
TYR"=>1,"ILELYS"=>1,"LYSILE"=>1,"ILEARG"=>1,"ARGILE"=>1,"ILEHIS"=>1,"HISILE"=>1,"METME
T"=>1,"METPHE"=>1,"PHEMET"=>1,"METTRP"=>1,"TRPMET"=>1,"METPRO"=>1,"PROMET"=>1,"METSER"
=>1,"SERMET"=>1,"METTHR"=>1,"THRMET"=>1,"METCYS"=>1,"CYSMET"=>1,"METTYR"=>1,"TYRMET"=>
1,"METLYS"=>1,"LYSMET"=>1,"METARG"=>1,"ARGMET"=>1,"METHIS"=>1,"HISMET"=>1,"PHEPHE"=>1,
"PHETRP"=>1,"TRPPHE"=>1,"PHEPRO"=>1,"PROPHE"=>1,"PHESER"=>1,"SERPHE"=>1,"PHETHR"=>1,"T
HRPHE"=>1,"PHECYS"=>1,"CYSPHE"=>1,"PHETYR"=>1,"TYRPHE"=>1,"PHELYS"=>1,"LYSPHE"=>1,"PHE
ARG"=>1,"ARGPHE"=>1,"PHEHIS"=>1,"HISPHE"=>1,"TRPTRP"=>1,"TRPPRO"=>1,"PROTRP"=>1,"TRPSE
R"=>1,"SERTRP"=>1,"TRPTHR"=>1,"THRTRP"=>1,"TRPCYS"=>1,"CYSTRP"=>1,"TRPTYR"=>1,"TYRTRP"
```

```
=>1,"TRPLYS"=>1,"LYSTRP"=>1,"TRPARG"=>1,"ARGTRP"=>1,"TRPHIS"=>1,"HISTRP"=>1,"PROPRO"=>
1,"PROSER"=>1,"SERPRO"=>1,"PROTHR"=>1,"THRPRO"=>1,"PROCYS"=>1,"CYSPRO"=>1,"PROTYR"=>1,
"TYRPRO"=>1,"PROLYS"=>1,"LYSPRO"=>1,"PROARG"=>1,"ARGPRO"=>1,"PROHIS"=>1,"HISPRO"=>1,"S
ERSER"=>1,"SERTHR"=>1,"THRSER"=>1,"SERCYS"=>1,"CYSSER"=>1,"SERTYR"=>1,"TYRSER"=>1,"SER
LYS"=>1,"LYSSER"=>1,"SERARG"=>1,"ARGSER"=>1,"SERHIS"=>1,"HISSER"=>1,"THRTHR"=>1,"THRCY
S"=>1,"CYSTHR"=>1,"THRTYR"=>1,"TYRTHR"=>1,"THRLYS"=>1,"LYSTHR"=>1,"THRARG"=>1,"ARGTHR"
=>1,"THRHIS"=>1,"HISTHR"=>1,"CYSCYS"=>1,"CYSTYR"=>1,"TYRCYS"=>1,"CYSLYS"=>1,"LYSCYS"=>
1,"CYSARG"=>1,"ARGCYS"=>1,"CYSHIS"=>1,"HISCYS"=>1,"TYRTYR"=>1,"TYRLYS"=>1,"LYSTYR"=>1,
"TYRARG"=>1,"ARGTYR"=>1,"TYRHIS"=>1,"HISTYR"=>1,"LYSLYS"=>1,"LYSARG"=>1,"ARGLYS"=>1,"L
YSHIS"=>1,"HISLYS"=>1,"ARGARG"=>1,"ARGHIS"=>1,"HISARG"=>1,"HISHIS"=>1,"ARGASP"=>1,"ASP
ARG"=>1,"ARGGLU"=>1,"GLUARG"=>1,"ARGASN"=>1,"ASNARG"=>1,"ARGGLN"=>1,"GLNARG"=>1,"LYSAS
P"=>1,"ASPLYS"=>1,"LYSGLN"=>1,"GLNLYS"=>1,"LYSASN"=>1,"ASNLYS"=>1,"LYSGLU"=>1,"GLULYS"
=>1,"ASPASP"=>1,"ASPGLU"=>1,"GLUASP"=>1,"ASPASN"=>1,"ASNASP"=>1,"ASPGLN"=>1,"GLNASP"=>
1,"ASPHIS"=>1,"HISASP"=>1,"ASPALA"=>1,"ALAASP"=>1,"ASPTYR"=>1,"TYRASP"=>1,"ASPTHR"=>1,
"THRASP"=>1,"ASPSER"=>1,"SERASP"=>1,"ASPPRO"=>1,"PROASP"=>1,"ASPGLY"=>1,"GLYASP"=>1,"G
LUGLU"=>1,"GLUASN"=>1,"ASNGLU"=>1,"GLUGLN"=>1,"GLNGLU"=>1,"GLUHIS"=>1,"HISGLU"=>1,"GLU
ALA"=>1,"ALAGLU"=>1,"GLUTYR"=>1,"TYRGLU"=>1,"GLUTHR"=>1,"THRGLU"=>1,"GLUSER"=>1,"SERGL
U"=>1,"GLUPRO"=>1,"PROGLU"=>1,"GLUGLY"=>1,"GLYGLU"=>1,"ASNASN"=>1,"ASNGLN"=>1,"GLNASN"
=>1,"ASNHIS"=>1,"HISASN"=>1,"ASNALA"=>1,"ALAASN"=>1,"ASNTYR"=>1,"TYRASN"=>1,"ASNTHR"=>
1,"THRASN"=>1,"ASNSER"=>1,"SERASN"=>1,"ASNPRO"=>1,"PROASN"=>1,"ASNGLY"=>1,"GLYASN"=>1,
"GLNGLN"=>1,"GLNHIS"=>1,"HISGLN"=>1,"GLNALA"=>1,"ALAGLN"=>1,"GLNTYR"=>1,"TYRGLN"=>1,"G
LNSER"=>1,"SERGLN"=>1,"GLNPRO"=>1,"PROGLN"=>1,"GLNGLY"=>1,"GLYGLN"=>1,"VALASP"=>1,"ASP
VAL"=>1,"VALASN"=>1,"ASNVAL"=>1,"VALGLN"=>1,"GLNVAL"=>1,"VALGLU"=>1,"GLUVAL"=>1,"ILEAS
P"=>1,"ASPILE"=>1,"ILEGLN"=>1,"GLNILE"=>1,"ILEASN"=>1,"ASNILE"=>1,"ILEGLN"=>1,"GLNILE"
=>1,"LEUASP"=>1,"ASPLEU"=>1,"LEUGLN"=>1,"GLNLEU"=>1,"LEUASN"=>1,"ASNLEU"=>1,"LEUGLN"=>
1,"GLNLEU"=>1,"METASP"=>1,"ASPMET"=>1,"METGLU"=>1,"GLUMET"=>1,"METASN"=>1,"ASNMET"=>1,
"METGLN"=>1,"GLNMET"=>1,"PHEASP"=>1,"ASPPHE"=>1,"PHEGLU"=>1,"GLUPHE"=>1,"PHEASN"=>1,"A
SNPHE"=>1,"PHEGLN"=>1,"GLNPHE"=>1,"TRPASP"=>1,"ASPTRP"=>1,"TRPGLU"=>1,"GLUTRP"=>1,"TRP
ASN"=>1,"ASNTRP"=>1,"TRPGLN"=>1,"GLNTRP"=>1,"CYSASP"=>1,"ASPCYS"=>1,"CYSGLU"=>1,"GLUCY
S"=>1,"CYSASN"=>1,"ASNCYS"=>1,"CYSGLN"=>1,"GLNCYS"=>1);

#This array (%atoms) includes all the accepted polar atom interactions, based on
#electrostatic charge. As such, interactions between nitrogen or oxygen atoms are not
#accepted. However, interactions between sulfur atoms are accepted, since they form
#disulfide brigdes.

%atoms=("N O "=>1,"N OG "=>1,"N OG1 "=>1,"N SG "=>1,"N OH "=>1,"N OD1 "=>1,"N OD2
"=>1,"N OE1 "=>1,"N OE2 "=>1,"O N "=>1,"O NE1 "=>1,"O NE2 "=>1,"O NZ "=>1,"O NE
"=>1,"O NH1 "=>1,"O NH2 "=>1,"O ND1 "=>1,"NE1 O "=>1,"NE1 OG "=>1,"NE1 OG1 "=>1,"NE1
SG "=>1,"NE1 OH "=>1,"NE1 OD1 "=>1,"NE1 OD2 "=>1,"NE1 OE1 "=>1,"NE1 OE2 "=>1,"OG N
"=>1,"OG NE1 "=>1,"OG NE2 "=>1,"OG NZ "=>1,"OG NE "=>1,"OG NH1 "=>1,"OG NH2 "=>1,"OG
ND1 "=>1,"OG1 N "=>1, "OG1 NE1 "=>1,"OG1 NE2 "=>1,"OG1 NZ "=>1,"OG1 NE "=>1,"OG1 NH1
"=>1,"OG1 NH2 "=>1,"OG1 ND1 "=>1,"SG SG "=>1,"SG N "=>1,"SG NE1 "=>1,"SG NE2 "=>1,"SG
NZ "=>1,"SG NE "=>1,"SG NH1 "=>1,"SG NH2 "=>1,"SG ND1 "=>1,"OH N "=>1,"OH NE1 "=>1,"OH
NE2 "=>1,"OH NZ "=>1,"OH NE "=>1,"OH NH1 "=>1,"OH NH2 "=>1,"OH ND1 "=>1,"OD1 N
"=>1,"OD1 NE1 "=>1,"OD1 NE2 "=>1,"OD1 NZ "=>1,"OD1 NE "=>1,"OD1 NH1 "=>1,"OD1 NH2
"=>1,"OD1 ND1 "=>1,"OD2 N "=>1,"OD2 NE1 "=>1,"OD2 NE2 "=>1,"OD2 NZ "=>1,"OD2 NE
"=>1,"OD2 NH1 "=>1,"OD2 NH2 "=>1,"OD2 ND1 "=>1,"OE1 N "=>1,"OE1 NE1 "=>1,"OE1 NE2
"=>1,"OE1 NZ "=>1,"OE1 NE "=>1,"OE1 NH1 "=>1,"OE1 NH2 "=>1,"OE1 ND1 "=>1,"NE2 O
"=>1,"NE2 OG "=>1,"NE2 OG1 "=>1,"NE2 SG "=>1,"NE2 OH "=>1,"NE2 OD1 "=>1,"NE2 OD2
"=>1,"NE2 OE1 "=>1,"NE2 OE1 "=>1,"NZ O "=>1,"NZ OG "=>1,"NZ OG1 "=>1,"NZ SG "=>1,"NZ
OH "=>1,"NZ OD1 "=>1,"NZ OD2 "=>1,"NZ OE1 "=>1,"NZ OE2 "=>1,"NE O "=>1,"NE OG "=>1,"NE
OG1 "=>1,"NE SG "=>1,"NE OH "=>1,"NE OD1 "=>1,"NE OD2 "=>1,"NE OE1 "=>1,"NE OE2
"=>1,"NH1 O "=>1,"NH1 OG "=>1,"NH1 OG1 "=>1,"NH1 SG "=>1,"NH1 OH "=>1,"NH1 OD1
"=>1,"NH1 OD2 "=>1,"NH1 OE1 "=>1,"NH1 OE2 "=>1,"NH2 O "=>1,"NH2 OG "=>1,"NH2 OG1
```

```
"=>1,"NH2 SG "=>1,"NH2 OH "=>1,"NH2 OD1 "=>1,"NH2 OD2 "=>1,"NH2 OE1 "=>1,"NH2 OE2
"=>1,"ND1 O "=>1,"ND1 OG "=>1,"ND1 OG1 "=>1,"ND1 SG "=>1,"ND1 OH "=>1,"ND1 OD1
"=>1,"ND1 OD2 "=>1,"ND1 OE1 "=>1,"ND1 OE2 "=>1);

#---------COMMANDS---------#

#First, open the total_contacts_<4-letter-pdb-name>.txt file. Then, we are going to
#check whether the amino acid pair is acceptable, as well as the atoms participating
#in the interaction. The filtered contacts will be exported to a new file, called
#ptotal_contacts_<4-letter-pdb-name>.txt.
@ARGV==1||die"Please enter a total_contacts file";
open(INPUT, $ARGV[0])||die"Cannot open total_contacts file.";
$ARGV[0]=~/(total_contacts\w+.txt)/;
open(OUTPUT, ">>p$1")||die"Cannot export data.";

#We are going to save the name of each amino acid and atom of every line. If the amino
#acid pair is acceptable, we are going to evaluate if the atoms are accepatable. If
#the atoms are hydrophobic, the amino acid pair is instantly accepted. However, if the
#atoms are polar, we need to verify that the contact is electrostaticallt favourable.
while($line=<INPUT>)
{
    chomp $line;
    if($line=~/\[PDB FILE\]\s\S+/)
    {
        print OUTPUT "\n$line\n";}
    if($line=~/^(\w+)\s(\w+)\s\w+\s\d+\s\-\s(\w+)\s(\w+)\s\w+\s\d+/)
    {
        $fat=$1;$faa=$2;$sat=$3;$saa=$4;
        $faa=~/(\w\w\w)$/;$faa=$1; #We need “MET”, not “AMET”
        $saa=~/(\w\w\w)$/;$saa=$1; #We need “LYS”, not “BLYS”
        if($pairs{$faa.$saa}==1)
        {
            if($aminos{$faa."H"}=~/$fat / && $aminos{$saa."H"}=~/$sat /)
            {
                print OUTPUT "$line\n";}
            elsif($aminos{$faa."P"}=~/$fat / && $aminos{$saa."P"}=~/$sat /)
            {
                if($atoms{"$fat $sat "}==1)
                {
                    print OUTPUT "$line\n";}}}}}
close(INPUT);
close(OUTPUT);
```

clusters.pl

```
use warnings;

#This script (clusters.pl) creates a list of all the contacts of all amino acids, in a
#more managable form for clusters2_1.pl to handle. Data are imported from any contacts
#file (species files are also contact files, they just include more than one
#structures) and exported in 20 files, one for each amino acid.

#---------COMMANDS---------#
@ARGV==1||die"Please enter total_contacts.txt .";
open(INPUT, $ARGV[0])||die"Cannot open file.";
open(ALA, ">>ala.dat")||die"Cannot export alanines.";
open(ARG, ">>arg.dat")||die"Cannot export arginines.";
open(ASP, ">>asp.dat")||die"Cannot export aspartic acids.";
open(ASN, ">>asn.dat")||die"Cannot export asparagines.";
open(CYS, ">>cys.dat")||die"Cannot export cysteins.";
open(GLU, ">>glu.dat")||die"Cannot export glutamic acids.";
open(GLN, ">>gln.dat")||die"Cannot export glutamines.";
open(GLY, ">>gly.dat")||die"Cannot exprot glycines.";
open(HIS, ">>his.dat")||die"Cannot export histidines";
open(ILE, ">>ile.dat")||die"Cannot export isoleucines.";
open(LEU, ">>leu.dat")||die"Cannot export leucines.";
open(LYS, ">>lys.dat")||die"Cannot export lysines.";
open(MET, ">>met.dat")||die"Cannot export metheionines.";
open(PHE, ">>phe.dat")||die"Cannot export phenylalanines.";
open(PRO, ">>pro.dat")||die"Cannot export prolines.";
open(SER, ">>ser.dat")||die"Cannot export serines.";
open(THR, ">>thr.dat")||die"Cannot export threonines.";
open(TRP, ">>trp.dat")||die"Cannot export tryptophans.";
open(TYR, ">>tyr.dat")||die"Cannot export tyrosines.";
open(VAL, ">>val.dat")||die"Cannot export valines.";
$i=0;
@tempaa=();
@tempan=();
@tempac=();
$tempaa[0]="Error";

#In any contacts file, the contacts of an amino acid are separated from contacts of
#another amino acid through blank lines. The format clusters2_1.pl needs is a display
#of all contacts of an amino acid in a single line.
while($line=<INPUT>)
{
    #If the line you read is the protein name, keep it. We will need it later.
    if($line=~/^\[PDB FILE\]\s(\S+)/)
    {
        $proteinname=$1;}
```

```
#If the line you read has information about a contact, save the names, number and
#chain of the amino contacts included. The first amino acid is the amino acid
#whose contacts are grouped. It is important to maintain any variance to the first
#amino contact name, since it may provide different groups of contacts, based on
#distance.
elsif($line=~/^\w+\s(\w+)\s(\w+)\s(\d+)\s\-\s\w+\s(\w+)\s(\w+)\s(\d+)/)
{
    $faa=$1; #AMET
    $fac=$2; #A
    $fan=$3; #1
    $saa=$4; #LEU
    $sac=$5; #B
    $san=$6; #3
    #If the name of the first amino acid already exists, we need to save just the
    #second amino acid details. Make sure that the second amino acid is not
    #already saved in the temporary arrays.
    if($tempaa[0] eq $faa)
    {
        if($saa!~/$tempaa[$i-1]$/ && $tempan[$i-1]!=$san)
        {
            $4=~/\w?(\w\w\w)$/;
            $tempaa[$i]=$1; #MET(not AMET)
            $tempan[$i]=$san; #38
            $tempac[$i]=$sac; #B
            $i++;}}
    #Your first amino acid is unlikely to be named Error. But this program need a
    #way to somehow begin saving contact data in the temporary arrays. This elsif
    #should be executed only at the beginning of each contacts file.
    elsif($tempaa[0] eq "Error")
    {
        $tempaa[$i]=$faa; #AMET(not MET)
        $tempan[$i]=$fan; #1
        $tempac[$i]=$fac; #A
        $i++;
        $4=~/\w?(\w\w\w)$/;
        $tempaa[$i]=$1; #LEU(not ALEU)
        $tempan[$i]=$san; #3
        $tempac[$i]=$sac; #B
        $i++;}
    #If you have found a blank line, then you have finished saving all the details
    #for a specific amino acid. You need to export the data to the respective
    #amino acid file and then reset the arrays to start over with the next amino
    #acid.
    else
    {
```

```
$tempaa[0]=~/\w?(\w\w\w)$/;
$tempaa[0]=$1;
$maxi=$i;
if($tempaa[0] eq "ALA")
{
    print ALA $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print ALA " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print ALA "\n";}
elsif($tempaa[0] eq "ARG")
{
    print ARG $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print ARG " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print ARG "\n";}
elsif($tempaa[0] eq "ASN")
{
    print ASN $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print ASN " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print ASN "\n";}
elsif($tempaa[0] eq "ASP")
{
    print ASP $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print ASP " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print ASP "\n";}
elsif($tempaa[0] eq "CYS")
{
    print CYS $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print CYS " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print CYS "\n";}
elsif($tempaa[0] eq "GLU")
{
    print GLU $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print GLU " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
```

```
    print GLU "\n";}
elsif($tempaa[0] eq "GLN")
{
    print GLN $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print GLN " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print GLN "\n";}
elsif($tempaa[0] eq "GLY")
{
    print GLY $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print GLY " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print GLY "\n";}
elsif($tempaa[0] eq "HIS")
{
    print HIS $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print HIS " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print HIS "\n";}
elsif($tempaa[0] eq "ILE")
{
    print ILE $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print ILE " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print ILE "\n";}
elsif($tempaa[0] eq "LEU")
{
    print LEU $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print LEU " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print LEU "\n";}
elsif($tempaa[0] eq "LYS")
{
    print LYS $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print LYS " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print LYS "\n";}
elsif($tempaa[0] eq "MET")
```

```
{
    print MET $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print MET " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print MET "\n";}
elsif($tempaa[0] eq "PHE")
{
    print PHE $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print PHE " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print PHE "\n";}
elsif($tempaa[0] eq "PRO")
{
    print PRO $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print PRO " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print PRO "\n";}
elsif($tempaa[0] eq "SER")
{
    print SER $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print SER " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print SER "\n";}
elsif($tempaa[0] eq "THR")
{
    print THR $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print THR " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print THR "\n";}
elsif($tempaa[0] eq "TRP")
{
    print TRP $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
    for($i=1;$i<$maxi;$i++)
    {
        print TRP " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
    print TRP "\n";}
elsif($tempaa[0] eq "TYR")
{
    print TYR $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
```

```
            for($i=1;$i<$maxi;$i++)
            {
                print TYR " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
            print TYR "\n";}
        elsif($tempaa[0] eq "VAL")
        {
            print VAL $proteinname, " ", $tempaa[0].$tempan[0].$tempac[0];
            for($i=1;$i<$maxi;$i++)
            {
                print VAL " ", $tempaa[$i].$tempan[$i].$tempac[$i];}
            print VAL "\n";}
        #Reset the arrays to start over.
        $i=0;
        $tempaa[$i]=$faa;
        $tempan[$i]=$fan;
        $tempac[$i]=$fac;
        $i++;
        $saa=~/\w?(\w\w\w)$/;
        $tempaa[$i]=$1;
        $tempan[$i]=$san;
        $tempac[$i]=$sac;
        $i++;}}
#If you found a no contacts line, then just export it to the respective amino
#acids file.
elsif($line=~/^(\w+)\s(\w+)\s(\d+)\s\-\sNo contacts/)
{
    $no=$3;
    $chain=$2;
    $name=$1;
    $name=~/\w?(\w\w\w)$/;
    if($1 eq "ALA")
    {
        print ALA $proteinname, " ", $1.$no.$chain, " No contacts\n";}
    elsif($1 eq "ARG")
    {
        print ARG $proteinname, " ", $1.$no.$chain, " No contacts\n";}
    elsif($1 eq "ASN")
    {
        print ASN $proteinname, " ", $1.$no.$chain, " No contacts\n";}
    elsif($1 eq "ASP")
    {
        print ASP $proteinname, " ", $1.$no.$chain, " No contacts\n";}
    elsif($1 eq "CYS")
    {
```

```
    print CYS $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "GLU")
{
    print GLU $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "GLN")
{
    print GLN $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "GLY")
{
    print GLY $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "HIS")
{
    print HIS $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "ILE")
{
    print ILE $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "LEU")
{
    print LEU $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "LYS")
{
    print LYS $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "MET")
{
    print MET $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "PHE")
{
    print PHE $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "PRO")
{
    print PRO $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "SER")
{
    print SER $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "THR")
{
    print THR $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "TRP")
{
    print TRP $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "TYR")
{
    print TYR $proteinname, " ", $1.$no.$chain, " No contacts\n";}
elsif($1 eq "VAL")
```

```
        {
            print VAL $proteinname, " ", $1.$no.$chain, " No contacts\n";}}}
close(INPUT);
close(ALA);close(ARG);close(ASN);close(ASP);close(CYS);close(GLU);
close(GLN);close(GLY);close(HIS);close(ILE);close(LEU);close(LYS);
close(MET);close(PHE);close(PRO);close(SER);close(THR);close(TRP);
close(TYR);close(VAL);
```

clusters2_1.pl

```
use warnings;

#This script (clusters2_1.pl) calculates the percantage of amino acid interactions for
#each amino acid. It receives as input the aa.dat files exported by clusters.pl and
#exports the preferredaa.dat files, which contain a list of the logarithmic percentage
#of each amino acid (including no contacts percentage) interaction for each amino
#acid. This script also calculates the average number of contacts for each amino acid,
#along with their respective standard deviation.

#---------HASH ARRAYS---------#

#This array (%names) contains all the full names for the amino acids. When exporting
#the final results to the preferredaa.dat files, we need the full names for the
#visualisation of our data through plots.r.
%names=("ala"=>"Alanine","arg"=>"Arginine","asn"=>"Asparagine","asp"=>"Aspartic_acid",
"cys"=>"Cystein","gln"=>"Glutamine","glu"=>"Glutamic_acid","gly"=>"Glycine","his"=>"Hi
stidine","ile"=>"Isoleucine","leu"=>"Leucine","lys"=>"Lysine","met"=>"Metheionine","ph
e"=>"Phenylalanine","pro"=>"Proline","ser"=>"Serine","thr"=>"Threonine","trp"=>"Trypto
phane","tyr"=>"Tyrosine","val"=>"Valine");

#---------COMMANDS---------#
@ARGV==1||die"Please enter a aa.dat file.\n";
open(INPUT, $ARGV[0])||die"Cannot open aa.dat file.\n";
$ARGV[0]=~/(\w\w\w\.dat)/;
$filename=$1;

#For calculating the interaction percantages, the $protein* variables and @proteins
#array are used, exporting to clusteraa.dat. For calculating the average number of
#contacts and standard deviation, the $amino* variables and @aminos array are used,
#exporting to dataaa.dat.
open(OUTPUT, ">>cluster$filename")||die"Cannot export cluster.\n";
open(DATA, ">>data$filename")||die"Cannot export data\n";
@proteins=();
$protein=0;
$proteincount=0;
$proteins[$protein]="Error";
@aminos=();
$aminocount=0;
$amino=0;

#Open e.g. pro.dat, and save the protein name and the aa names.
while($line=<INPUT>)
{
    if($line=~/(\w+)\.ent\s\w\w\w\d+\w+\s(\w\w\w)\d+\w+/)
    {
        #See if the name you got already exists
        $proteini=$protein;
        while($protein>=0)
        {
            if($proteins[$protein] eq $1)
```

```
                {
                        $proteincount++;}
                $protein--;}
            #But, if there isn't, add it
            if($proteincount==0)
            {
                $proteins[$proteini]=$1;
                $proteini++;
                $proteins[$proteini]="Error";}
            $proteincount=0;
            $protein=$proteini;
            print OUTPUT "$2 ";
            $amino++;
            #For more than one contact, you need to search the remaining line again.
            while($'=~/(\w\w\w)\d+\w+/)
            {
                print OUTPUT "$1 ";
                $amino++;}
            print OUTPUT "\n";
            $aminos[$aminocount]=$amino;
            $aminocount++;}
        #What if you line doesn't include any contacts?
        elsif($line=~/^.*\.ent\s\w\w\w\d+\w+\sNo contacts/)
        {
            print OUTPUT "No contacts\n";}
        $amino=0;}
    #Calculate the average value of contacts.
    $aminofinal=$aminocount;
    $aminosum=0;
    for($aminocount=0;$aminocount<$aminofinal;$aminocount++)
    {
        $aminosum=$aminosum+$aminos[$aminocount];}
    $aminoaverage=$aminosum/($aminofinal-1);
    #Also, calculate the standard deviation.
    $aminos2=0;
    for($aminocount=0;$aminocount<$aminofinal;$aminocount++)
    {
        $aminos2=$aminos2+($aminos[$aminocount]-$aminoaverage)*($aminos[$aminocount]-
        $aminoaverage);}
    $amino_sd=sqrt($aminos2/($aminofinal-1));
    #Print: In the ala.dat file, I found 16534 proteins that give for n alanines (you can
    deduce that n by the no of lines this file has) 64.8 contacts average with +- 32.064
    standard deviation [In short: ala.dat 16534 64.8 +- 32.064].
```

```
print DATA "$filename $proteini @aminos";
print "$filename $proteini $aminoaverage +- $amino_sd\n";
close(INPUT);
close(DATA);
close(OUTPUT);
#Open the clusteraa.dat file and let's start processing.
open(INPUT, "cluster$filename")||die"Cannot open cluster$ARGV[0].\n";
open(OUTPUT, ">>preferred$filename")||die"Cannot export data.\n";
$filename=~/^(\w\w\w)/;
$name=$1;
#Load everything on an array.
@array=();
$i=0;
while($line=<INPUT>)
{
    $array[$i]=$line;
    $i++;}
$imax=$i;
$nocontacts=0;
$ala=0;$arg=0;$asn=0;$asp=0;$cys=0;$gln=0;$glu=0;$gly=0;$his=0;$ile=0;
$leu=0;$lys=0;$met=0;$phe=0;$pro=0;$ser=0;$thr=0;$trp=0;$tyr=0;$val=0;
#While searching the array
for($i=0;$i<$imax;$i++)
{
    $icounted=$i;
    #Count the number of 'No contacts' lines
    if($array[$i]=~/^No\scontacts/)
    {
        $nocontacts++;}
    elsif($array[$i]=~/(\w+)\s/)
    {
        if($1 eq 'ALA')
        {$ala++;}
        elsif($1 eq 'ARG')
        {$arg++;}
        elsif($1 eq 'ASN')
        {$asn++;}
        elsif($1 eq 'ASP')
        {$asp++;}
        elsif($1 eq 'CYS')
        {$cys++;}
        elsif($1 eq 'GLN')
        {$gln++;}
```

```
elsif($1 eq 'GLU')
{$glu++;}
elsif($1 eq 'GLY')
{$gly++;}
elsif($1 eq 'HIS')
{$his++;}
elsif($1 eq 'ILE')
{$ile++;}
elsif($1 eq 'LEU')
{$leu++;}
elsif($1 eq 'LYS')
{$lys++;}
elsif($1 eq 'MET')
{$met++;}
elsif($1 eq 'PHE')
{$phe++;}
elsif($1 eq 'PRO')
{$pro++;}
elsif($1 eq 'SER')
{$ser++;}
elsif($1 eq 'THR')
{$thr++;}
elsif($1 eq 'TRP')
{$trp++;}
elsif($1 eq 'TYR')
{$tyr++;}
elsif($1 eq 'VAL')
{$val++;}
#(Do this recursively for more than one aa)
while($'=~/(\w+)\s/)
{
       if($1 eq 'ALA' && $icounted!=$i)
       {$ala++;}
        elsif($1 eq 'ARG' && $icounted!=$i)
       {$arg++;}
       elsif($1 eq 'ASN' && $icounted!=$i)
       {$asn++;}
        elsif($1 eq 'ASP' && $icounted!=$i)
       {$asp++;}
         elsif($1 eq 'CYS' && $icounted!=$i)
        {$cys++;}
       elsif($1 eq 'GLN' && $icounted!=$i)
          {$gln++;}
         elsif($1 eq 'GLU' && $icounted!=$i)
```

```
                    {$glu++;}
                       elsif($1 eq 'GLY' && $icounted!=$i)
                     {$gly++;}
                    elsif($1 eq 'HIS' && $icounted!=$i)
                       {$his++;}
                     elsif($1 eq 'ILE' && $icounted!=$i)
                    {$ile++;}
                       elsif($1 eq 'LEU' && $icounted!=$i)
                     {$leu++;}
                    elsif($1 eq 'LYS' && $icounted!=$i)
                       {$lys++;}
                     elsif($1 eq 'MET' && $icounted!=$i)
                    {$met++;}
                       elsif($1 eq 'PHE' && $icounted!=$i)
                     {$phe++;}
                    elsif($1 eq 'PRO' && $icounted!=$i)
                       {$pro++;}
                     elsif($1 eq 'SER' && $icounted!=$i)
                    {$ser++;}
                       elsif($1 eq 'THR' && $icounted!=$i)
                     {$thr++;}
                    elsif($1 eq 'TRP' && $icounted!=$i)
                       {$trp++;}
                     elsif($1 eq 'TYR' && $icounted!=$i)
                    {$tyr++;}
                       elsif($1 eq 'VAL' && $icounted!=$i)
                     {$val++;}}}}
#Time to export our results to preferredaa.dat, to be visualised by plots.r.
print OUTPUT "1 Contact Percentage\n";
print OUTPUT $names{"$name"}, " ", $names{"$name"}, " ", log(100), "\n";
if($nocontacts!=0)
{print OUTPUT "None None ", (log(($nocontacts/$imax)*100)), "\n";}
if($ala!=0)
{print OUTPUT "ALA ALA ", (log(($ala*100)/$imax)), "\n";}
if($arg!=0)
{print OUTPUT "ARG ARG ", (log(($arg*100)/$imax)), "\n";}
if($asn!=0)
{print OUTPUT "ASN ASN ", (log(($asn*100)/$imax)), "\n";}
if($asp!=0)
{print OUTPUT "ASP ASP ", (log(($asp*100)/$imax)), "\n";}
if($cys!=0)
{print OUTPUT "CYS CYS ", (log(($cys*100)/$imax)), "\n";}
if($gln!=0)
{print OUTPUT "GLN GLN ", (log(($gln*100)/$imax)), "\n";}
```

```
if($glu!=0)
{print OUTPUT "GLU GLU ", (log(($glu*100)/$imax)), "\n";}
if($gly!=0)
{print OUTPUT "GLY GLY ", (log(($gly*100)/$imax)), "\n";}
if($his!=0)
{print OUTPUT "HIS HIS ", (log(($his*100)/$imax)), "\n";}
if($ile!=0)
{print OUTPUT "ILE ILE ", (log(($ile*100)/$imax)), "\n";}
if($leu!=0)
{print OUTPUT "LEU LEU ", (log(($leu*100)/$imax)), "\n";}
if($lys!=0)
{print OUTPUT "LYS LYS ", (log(($lys*100)/$imax)), "\n";}
if($met!=0)
{print OUTPUT "MET MET ", (log(($met*100)/$imax)), "\n";}
if($phe!=0)
{print OUTPUT "PHE PHE ", (log(($phe*100)/$imax)), "\n";}
if($pro!=0)
{print OUTPUT "PRO PRO ", (log(($pro*100)/$imax)), "\n";}
if($ser!=0)
{print OUTPUT "SER SER ", (log(($ser*100)/$imax)), "\n";}
if($thr!=0)
{print OUTPUT "THR THR ", (log(($thr*100)/$imax)), "\n";}
if($trp!=0)
{print OUTPUT "TRP TRP ", (log(($trp*100)/$imax)), "\n";}
if($tyr!=0)
{print OUTPUT "TYR TYR ", (log(($tyr*100)/$imax)), "\n";}
if($val!=0)
{print OUTPUT "VAL VAL ", (log(($val*100)/$imax)), "\n";}
close(OUTPUT);
close(INPUT);
#Remove the clusteraa.dat file, since it no longer needed.
system("rm cluster$filename");
```

plots.r

```r
# Libraries
library(packcircles)
library(ggplot2)

# Generate the layout. This function return a dataframe with one line per bubble.
# It gives its center (x and y) and its radius, proportional of the value
packing <- circleProgressiveLayout(preferredval$Percentage, sizetype='area')

# We can add these packing information to the initial data frame
data <- cbind(preferredval, packing)

# Check that radius is proportional to value. We don't want a linear relationship,
since it is the AREA that must be proportional to the value
plot(data$radius, data$Percentage)

# The next step is to go from one center + a radius to the coordinates of a circle
that is drawn by a multitude of straight lines.
dat.gg <- circleLayoutVertices(packing, npoints=50)

# Make the plot
ggplot() +

  # Make the bubbles
  geom_polygon(data = dat.gg, aes(x, y, group = id), colour = "white", alpha = 0.1) +

  # Add text in the center of each bubble + control its size
  geom_text(data = data, aes(x, y, size=Percentage, label = Contact)) +
  scale_size_continuous(range = c(1,5)) +

  # General theme:
  theme_void() +
  theme(legend.position="bottom") +
  coord_equal()
```

Amino acid interaction percentages

Table 2 Amino acid interaction percentages for bacteria. Analysis performed at 5Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,34	2,09	1,67	1,78	2,1	1,85	1,7	1,92	1,71	2,04	1,88	1,98	1,66	2,04	1,93	1,87	1,72	1,69	1,3	2,01
Arginine	1,6	2,01	1,8	2,28	1,16	1,76	2,24	1,78	1,32	1,67	1,82	1,19	1,37	1,61	2,02	1,69	1,91	1,34	1,42	1,72
Asparagine	1,26	1,4	2,09	1,81	2,03	1,67	1,82	1,5	1,42	1,38	0,99	1,55	1,63	1,32	1,59	1,81	1,94	2,06	1,68	1,52
Aspartic acid	1,68	2,08	1,81	1,79	0,47	2,11	1,93	1,8	2,34	1,46	1,57	1,96	1,16	1,87	1,7	2,23	2,08	1,88	1,79	1,59
Cystein	0,04	0,06	0,03	-0,92	0,98	0,75	-0,12	-0,64	0,39	0,15	0,06	-0,49	0,69	0,52	-0,48	0	0,12	0,38	-0,27	0,29
Glutamine	1,14	1,49	1,26	1,39	1,45	1,77	1,38	1,31	0,95	1,04	1,08	1,23	0,95	1,48	1,24	1,9	1,52	1,22	1,45	1,22
Glutamic acid	1,44	2,23	1,85	1,68	1,1	1,53	1,75	1,68	1,89	1,54	1,65	2,09	1,23	1,56	1,82	1,72	1,53	1,5	1,37	1,73
Glycine	1,8	1,71	1,84	1,77	1,77	1,46	1,77	1,96	1,67	1,65	1,51	1,79	1,86	1,62	2,06	1,57	1,73	1,55	1,56	1,31
Histidine	0,63	0,7	1,27	1,22	0,91	0,3	1,36	0,99	1,28	0,77	1,38	0,42	0,83	1,2	0,75	0,99	1,33	1,53	1,51	1,04
Isoleucine	1,77	1,36	1,81	1,34	2,61	1,77	1,34	1,65	1,28	2,02	2,03	1,82	2,38	2,01	1,57	1,52	1,35	1,75	1,92	1,76
Leucine	2,35	2,22	2,08	2,16	2,31	2,31	2,1	2,4	2,2	2,48	2,59	2,18	1,79	2,29	2,01	1,95	2,22	1,99	2,55	2,52
Lysine	1,3	1,38	1,44	1,7	1,27	0,75	1,66	1,55	1,54	1,61	1,39	1,84	1,54	1,02	1,3	1,57	1,29	1,71	1,63	1,26
Methionine	1,08	0,67	0,78	1,07	0,06	0,52	0,86	0,42	1,11	0,62	0,95	1,05	1,65	1,34	1,28	0,33	0,83	1,07	0,9	1,03
Phenylalanine	1,8	1,86	1,64	1,19	1,67	1,55	1,37	1,84	1,11	1,89	1,87	1,61	1,86	1,88	1,59	1,55	1,54	1,93	1,67	1,92
Proline	1,49	1,2	1,04	0,95	0,22	1,48	1,5	1,14	0,88	1,08	1,24	1,48	1,14	1,02	1,04	1,32	1,43	1,55	1,15	1,11
Serine	1,66	1,55	2	2,04	1,8	1,78	1,81	1,64	1,99	1,54	1,42	1,42	1,59	1,24	1,96	1,87	1,78	1,4	1,98	1,41
Threonine	1,86	1,41	1,55	1,69	1,16	1,93	2,08	1,82	1,93	1,69	1,59	1,4	1,84	1,78	1,62	2,02	1,81	1,94	1,41	1,54
Tryptophane	0,77	0,89	0,74	1,04	-0,63	1,04	0,46	0,92	1,06	0,89	0,97	1,12	1,35	1,19	1,07	0,78	0,86	0,95	1,24	0,86
Tyrosine	1,57	1,36	1,58	1,49	2,1	1,55	1,11	1,58	1,93	1,76	1,48	2,01	1,93	1,83	1,8	1,46	1,27	1,6	1,8	1,58
Valine	1,99	1,83	1,8	1,81	2,12	1,92	1,6	1,92	1,92	2,14	2,15	1,76	2,02	1,73	1,87	1,64	1,95	1,8	1,81	2,32

Table 3 Amino acid interaction percentages for bacteria. Analysis performed at 7Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,44	2,41	1,93	1,95	2,51	2,3	2,08	2,24	2,02	2,37	2,34	2,16	2,24	2,18	2,12	2,14	2,07	2,53	1,93	2,33
Arginine	1,55	1,86	1,36	1,79	1,87	1,43	2,06	1,73	0,96	1,16	1,4	1,48	1,44	1,2	1,9	1,43	1,45	1,54	1,62	1,57
Asparagine	1,34	1,43	1,73	1,73	1,73	1,54	1,64	1,45	1,56	1,2	1,33	1,51	1,77	1,12	1,54	1,64	1,92	1,51	1,22	1,26
Aspartic acid	1,64	2,04	1,97	1,68	1	2,13	1,95	1,79	2	1,33	1,67	2,13	1,16	1,71	1,8	2,17	1,94	1,85	1,91	1,57
Cystein	-0,24	0,23	0,52	-0,8	1	-0,62	0,36	-0,03	0,83	0,31	0,06	-0,06	0,5	0,51	0,07	-0,08	0,02	0,02	-0,79	-0,17
Glutamine	1,17	1,48	1,12	1,29	1,21	1,45	1,38	1,08	0,76	0,97	0,92	0,96	0,21	1,06	1,07	1,43	1,36	1	1,19	0,85
Glutamic acid	1,7	2,06	1,78	1,54	0,74	1,64	1,52	1,8	1,61	1,48	1,53	1,99	1,34	1,69	1,41	1,54	1,39	0,87	1,35	1,48
Glycine	2,23	1,91	2,05	1,98	2,07	1,71	1,95	2,12	2,03	2,05	2,06	1,99	1,88	2,05	2,48	2,06	1,76	2,1	1,93	2,02
Histidine	0,6	0,69	0,75	1,42	1,33	0,93	1,38	0,97	0,93	0,95	1,24	-0,11	0,86	0,8	0,84	1,02	1,14	1,73	1,22	0,77
Isoleucine	1,55	1,21	1,85	1,28	2,24	1,72	1,61	1,59	1,63	1,82	1,91	1,75	2,11	1,99	1,54	1,27	1,82	1,54	1,84	1,99
Leucine	2,29	1,91	2,2	2,3	2,12	2,3	2,01	2,27	2,42	2,38	2,35	2,13	2,18	2,42	1,93	2,12	2,2	2,03	2,27	2,36
Lysine	1,31	1,13	1,32	1,5	0,38	1,19	1,58	1,4	1,49	1,47	1,08	1,65	1,31	1,17	1,54	1,55	1,32	0,77	1,3	1,05
Methionine	0,73	0,84	0,81	0,88	0,38	0,89	0,61	0,32	1,14	0,87	0,83	0,87	1,88	1,6	1,15	0,99	0,92	1,32	1,1	0,69
Phenylalanine	1,5	1,71	1,43	1,42	1,61	1,36	1,18	1,81	1,54	1,73	1,64	1,49	1,9	1,82	1,56	1,56	1,35	1,83	1,62	1,68
Proline	1,64	1,16	1,35	1,56	1,14	1,55	1,61	1,23	1,51	1,41	1,5	1,55	0,86	1,29	1,41	1,59	1,72	1,86	1,27	1,47
Serine	1,62	1,55	2,21	2,05	1,99	1,81	1,82	1,82	1,66	1,57	1,7	1,45	1,53	1,53	1,94	1,84	1,82	1,56	2,01	1,84
Threonine	1,83	1,67	1,76	1,97	1,38	1,69	1,88	1,75	1,86	1,52	1,79	1,89	1,75	1,82	1,58	2,06	2,05	1,51	1,89	1,56
Tryptophane	0,34	0,81	0,24	0,64	0,23	0,64	0,47	0,75	0,59	0,67	0,21	0,92	0,43	0,73	0,72	0,33	0,67	0,39	0,35	0,63
Tyrosine	1,34	1,43	1,65	1,13	1,33	1,24	0,96	1,25	1,52	1,87	1,23	1,67	1,87	1,42	1,2	1,3	1,13	1,12	1,84	1,4
Valine	2,04	2,04	1,62	1,89	2,05	2,09	1,9	1,91	1,95	2,2	2,16	1,68	1,94	1,74	1,85	1,57	1,82	2	1,93	2,27

Table 4 Amino acid interaction percentages for archaea. Analysis performed at 5Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,01	1,78	1,62	1,59	1,84	1,76	1,63	1,96	1,77	1,91	1,95	1,68	1,82	1,8	1,68	1,77	1,64	1,71	1,84	1,99
Arginine	1,62	2,05	1,75	2,16	1,89	1,64	2,04	1,86	1,73	1,66	1,7	1,65	1,85	1,64	1,85	1,78	1,85	2	1,73	1,66
Asparagine	1,17	1,3	1,92	1,58	1,03	1,54	1,49	1,27	1,43	1,21	1,16	1,44	1,23	1,26	1,55	1,49	1,48	1,45	1,27	1,26
Aspartic acid	1,66	2,02	2	1,99	1,46	2,01	1,88	1,64	1,81	1,56	1,6	2,12	1,54	1,6	1,69	2,03	2,04	1,72	1,58	1,5
Cystein	-0,13	-0,4	-0,29	-0,25	2,47	0,04	-0,47	0,07	0,5	-0,32	-0,46	-0,52	0,35	-0,14	-0,11	-0,25	-0,07	-0,41	-0,17	-0,16
Glutamine	0,54	0,74	0,79	0,78	0,93	1,3	0,69	0,59	0,77	0,57	0,53	0,69	0,77	0,67	0,87	0,76	0,74	0,79	0,84	0,54
Glutamic acid	1,77	2,39	2,09	1,84	1,46	2,15	2,3	1,67	2,07	1,8	1,87	2,36	1,81	1,8	1,91	2,03	2,02	1,79	1,91	1,7
Glycine	1,6	1,64	1,66	1,64	1,57	1,61	1,49	1,82	1,78	1,51	1,5	1,57	1,5	1,62	1,89	1,73	1,89	1,6	1,43	1,69
Histidine	0,67	0,47	0,75	0,85	0,77	0,91	0,61	0,87	1,39	0,49	0,58	0,37	0,52	0,71	0,61	0,84	0,89	0,53	0,55	0,51
Isoleucine	2,31	1,99	2,03	1,92	2,02	1,96	1,96	2,14	1,95	2,48	2,35	2,1	2,26	2,3	1,98	2,05	2,1	2,19	2,25	2,37
Leucine	2,45	2,21	2,07	2,02	2,16	2,28	2,08	2,42	2,15	2,45	2,51	2,21	2,31	2,34	2,16	2,17	2,11	2,16	2,37	2,39
Lysine	1,68	1,56	2,02	2,22	1,3	1,81	2,3	1,8	1,64	1,71	1,7	1,97	1,65	1,76	1,69	1,78	1,74	1,85	1,84	1,7
Methionine	0,92	0,6	0,67	0,79	0,84	0,77	0,73	0,86	0,8	0,88	0,79	0,7	1,45	0,97	0,76	0,69	0,81	1,17	0,83	0,97
Phenylalanine	1,64	1,56	1,5	1,23	1,46	1,42	1,29	1,58	1,47	1,58	1,65	1,39	1,69	1,86	1,65	1,46	1,44	1,67	1,6	1,56
Proline	1,3	1,29	1,34	1,35	1,25	1,22	1,4	1,29	1,53	1,23	1,28	1,26	1,5	1,42	1,55	1,43	1,25	1,47	1,36	1,15
Serine	1,51	1,5	1,57	1,87	1,54	1,55	1,77	1,43	1,56	1,47	1,49	1,64	1,54	1,54	1,69	1,86	1,68	1,59	1,57	1,44
Threonine	1,54	1,42	1,59	1,71	1,48	1,73	1,59	1,62	1,75	1,46	1,54	1,39	1,39	1,56	1,59	1,61	1,69	1,48	1,55	1,52
Tryptophane	0,09	0,51	0,26	0	0,23	0,62	0,11	0,25	0,59	0,1	0,34	0,23	0,13	0,41	0,66	0,01	0,23	0,82	0,4	0,21
Tyrosine	1,62	1,69	1,7	1,6	1,38	1,58	1,47	1,63	1,6	1,54	1,61	1,66	1,8	1,63	1,87	1,5	1,49	1,75	1,77	1,5
Valine	2,3	2,02	1,92	1,87	2,31	1,85	1,84	2,18	2	2,4	2,26	2,05	2,15	2,19	2,03	2,06	2,11	2	2,22	2,5

Table 5 Amino acid interaction percentages for archaea. Analysis performed at 7Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,36	2	1,94	1,85	2,11	2,05	1,83	2,17	2,09	2,29	2,3	1,85	2,2	2,1	2,02	1,96	2,01	2,16	2,11	2,32
Arginine	1,32	1,52	1,47	1,84	1,49	1,49	1,94	1,6	1,44	1,52	1,51	1,43	1,55	1,47	1,54	1,56	1,48	1,77	1,57	1,45
Asparagine	1	1,33	1,63	1,51	0,98	1,52	1,45	1,22	1,36	1,11	1	1,49	1,1	1,07	1,44	1,52	1,39	1,19	1,13	1,1
Aspartic acid	1,51	2,05	1,96	1,84	1,55	1,94	1,88	1,61	1,81	1,49	1,47	2,07	1,43	1,54	1,64	2,02	1,9	1,6	1,46	1,37
Cystein	-0,1	-0,42	-0,23	-0,25	2,46	-0,07	-0,41	0,01	0,44	-0,18	-0,2	-0,44	0,16	-0,22	-0,04	-0,11	0,07	-0,32	0,06	-0,08
Glutamine	0,33	0,58	0,63	0,68	0,28	0,76	0,67	0,49	0,54	0,38	0,36	0,46	0,7	0,46	0,66	0,5	0,48	0,48	0,78	0,37
Glutamic acid	1,64	2,3	2,19	1,83	1,23	2,08	2,11	1,74	1,93	1,74	1,76	2,45	1,66	1,69	1,94	1,94	1,86	1,67	1,76	1,62
Glycine	1,83	1,94	1,96	1,94	2,01	1,85	1,77	2,11	1,98	1,83	1,83	1,79	1,93	1,86	2,14	2	2,11	2,03	1,78	1,9
Histidine	0,35	0,4	0,48	0,7	0,83	0,78	0,49	0,6	1,34	0,44	0,37	0,37	0,44	0,72	0,57	0,75	0,78	0,52	0,4	0,38
Isoleucine	2,39	2,03	2,11	1,99	2,17	1,91	1,96	2,13	2,03	2,33	2,32	2,13	2,24	2,29	1,97	2,01	2,13	2,17	2,27	2,34
Leucine	2,33	2,21	2,05	2,04	2,01	2,19	2,03	2,32	2,16	2,33	2,45	2,13	2,29	2,39	2,17	2,06	2,07	2,09	2,37	2,29
Lysine	1,53	1,61	1,91	2,17	1,24	1,82	2,38	1,82	1,51	1,74	1,7	1,92	1,66	1,68	1,72	1,85	1,59	1,82	1,75	1,69
Methionine	0,8	0,6	0,59	0,89	0,89	0,55	0,71	0,81	0,77	0,82	0,76	0,57	1,12	0,9	0,8	0,69	0,77	0,96	0,86	0,85
Phenylalanine	1,33	1,39	1,39	1,2	1,43	1,37	1,19	1,36	1,35	1,38	1,52	1,27	1,44	1,64	1,43	1,3	1,38	1,72	1,48	1,44
Proline	1,55	1,55	1,52	1,7	1,36	1,52	1,62	1,5	1,66	1,49	1,47	1,41	1,64	1,73	1,76	1,62	1,6	1,64	1,62	1,48
Serine	1,59	1,62	1,76	1,92	1,5	1,65	1,75	1,63	1,61	1,56	1,5	1,69	1,47	1,53	1,72	1,89	1,72	1,53	1,62	1,52
Threonine	1,58	1,48	1,54	1,66	1,39	1,62	1,5	1,63	1,71	1,58	1,56	1,33	1,47	1,48	1,64	1,68	1,73	1,53	1,49	1,64
Tryptophane	-0,14	0,18	0,02	-0,28	-0,01	0,25	-0,29	0,01	0,37	-0,33	-0,05	-0,12	-0,27	0,15	-0,1	-0,02	-0,15	0,16	0,15	-0,11
Tyrosine	1,34	1,44	1,46	1,28	1,32	1,55	1,25	1,31	1,34	1,41	1,4	1,56	1,61	1,46	1,5	1,31	1,27	1,63	1,63	1,4
Valine	2,51	2,13	1,97	1,98	2,34	2,12	1,97	2,22	2,14	2,39	2,33	2,1	2,38	2,28	2,15	2,06	2,29	2,13	2,23	2,45

Table 6 Amino acid interaction percentages for eukarya. Analysis performed at 5Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	1,88	1,7	1,61	1,58	1,59	1,66	1,62	1,78	1,66	1,76	1,77	1,66	1,71	1,69	1,59	1,65	1,68	1,69	1,67	1,81
Arginine	1,66	2,05	1,75	2,02	1,57	1,72	1,94	1,78	1,66	1,54	1,62	1,55	1,61	1,66	1,82	1,7	1,72	1,59	1,65	1,53
Asparagine	1,47	1,44	1,99	1,7	1,22	1,63	1,62	1,46	1,61	1,35	1,31	1,6	1,31	1,42	1,54	1,63	1,65	1,51	1,5	1,33
Aspartic acid	1,69	1,95	1,99	1,97	1,47	1,83	1,77	1,71	1,88	1,47	1,51	2,04	1,52	1,54	1,64	1,96	1,92	1,62	1,61	1,53
Cystein	0,68	0,59	0,61	0,61	2,44	0,62	0,46	0,79	0,8	0,76	0,7	0,55	0,55	0,74	0,72	0,65	0,7	0,86	0,77	0,86
Glutamine	1,35	1,43	1,41	1,49	1,25	1,81	1,47	1,35	1,33	1,25	1,32	1,41	1,3	1,32	1,41	1,48	1,43	1,32	1,29	1,22
Glutamic acid	1,69	2,04	1,82	1,58	1,41	1,83	2,01	1,62	1,82	1,59	1,64	2,07	1,68	1,64	1,64	1,83	1,73	1,49	1,59	1,59
Glycine	1,55	1,55	1,58	1,61	1,41	1,42	1,5	1,77	1,56	1,43	1,43	1,5	1,46	1,47	1,75	1,69	1,64	1,62	1,5	1,6
Histidine	0,92	0,94	1,06	1,13	1,13	0,93	0,99	1,05	1,49	0,93	0,95	0,85	1,04	0,99	1,07	1	1,07	1,03	1	0,96
Isoleucine	1,84	1,7	1,67	1,53	1,87	1,62	1,59	1,79	1,74	2,19	1,99	1,73	1,93	1,96	1,66	1,65	1,75	1,86	1,9	2
Leucine	2,43	2,31	2,14	2,05	2,35	2,3	2,22	2,3	2,27	2,54	2,61	2,26	2,45	2,45	2,17	2,2	2,17	2,37	2,43	2,46
Lysine	1,57	1,47	1,75	2,02	1,37	1,7	1,97	1,61	1,58	1,54	1,56	1,85	1,56	1,53	1,59	1,7	1,65	1,59	1,64	1,51
Methionine	0,88	0,76	0,69	0,56	0,83	0,89	0,76	0,88	0,72	0,98	0,91	0,76	1,55	0,99	0,83	0,68	0,73	0,82	0,91	0,91
Phenylalanine	1,73	1,58	1,6	1,47	1,74	1,52	1,44	1,72	1,65	1,81	1,81	1,59	1,9	1,94	1,7	1,59	1,61	1,78	1,82	1,78
Proline	1,4	1,32	1,39	1,42	1,23	1,39	1,43	1,44	1,34	1,21	1,25	1,34	1,28	1,34	1,72	1,46	1,41	1,55	1,41	1,29
Serine	1,82	1,74	1,89	2	1,6	1,87	1,92	1,73	1,85	1,66	1,66	1,81	1,59	1,71	1,86	2,08	1,88	1,75	1,68	1,71
Threonine	1,67	1,56	1,72	1,76	1,56	1,78	1,83	1,67	1,64	1,64	1,59	1,65	1,63	1,6	1,63	1,66	1,84	1,53	1,64	1,63
Tryptophane	0,72	0,8	0,64	0,57	0,85	0,76	0,55	0,88	0,75	0,78	0,8	0,65	0,91	0,81	1,05	0,66	0,68	1,13	0,77	0,76
Tyrosine	1,49	1,54	1,5	1,59	1,52	1,5	1,38	1,58	1,59	1,57	1,54	1,59	1,55	1,67	1,82	1,45	1,44	1,68	1,81	1,52
Valine	1,99	1,85	1,74	1,71	1,99	1,84	1,75	1,95	1,89	2,11	2,11	1,83	2,05	2,04	1,82	1,8	1,96	2,07	1,98	2,25

Table 7 Amino acid interaction percentages for eukarya. Analysis performed at 7Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,2	1,93	1,85	1,81	1,91	1,86	1,85	2	1,96	2,08	2,08	1,89	1,98	2,01	1,89	1,91	1,94	1,96	1,95	2,08
Arginine	1,46	1,52	1,44	1,7	1,27	1,57	1,81	1,54	1,47	1,43	1,47	1,38	1,44	1,37	1,54	1,47	1,42	1,46	1,44	1,39
Asparagine	1,37	1,48	1,77	1,68	1,19	1,6	1,56	1,46	1,48	1,28	1,24	1,6	1,29	1,35	1,54	1,59	1,63	1,41	1,4	1,25
Aspartic acid	1,63	1,97	2	1,79	1,34	1,79	1,77	1,68	1,79	1,41	1,43	2,06	1,45	1,45	1,67	1,94	1,87	1,47	1,47	1,45
Cystein	0,78	0,71	0,71	0,73	2,55	0,68	0,59	0,82	0,93	0,87	0,82	0,6	0,86	0,83	0,81	0,81	0,71	0,91	0,9	0,94
Glutamine	1,2	1,33	1,36	1,44	0,96	1,52	1,48	1,29	1,25	1,13	1,21	1,33	1,18	1,16	1,32	1,34	1,31	1,21	1,11	1,1
Glutamic acid	1,67	2,04	1,82	1,57	1,27	1,85	1,74	1,61	1,7	1,58	1,6	2,11	1,6	1,55	1,66	1,78	1,75	1,52	1,5	1,57
Glycine	1,79	1,81	1,84	1,86	1,77	1,75	1,73	2,01	1,86	1,75	1,7	1,72	1,8	1,77	2,02	1,94	1,92	1,81	1,8	1,82
Histidine	0,81	0,85	0,94	1,13	0,95	0,85	0,87	0,92	1,19	0,83	0,82	0,78	0,92	0,91	0,96	0,92	0,92	0,87	0,94	0,86
Isoleucine	1,88	1,7	1,68	1,59	1,92	1,7	1,63	1,77	1,8	2,07	1,97	1,72	1,98	1,99	1,6	1,7	1,75	1,88	1,92	1,94
Leucine	2,32	2,28	2,1	2,1	2,28	2,26	2,16	2,26	2,29	2,45	2,53	2,21	2,39	2,44	2,19	2,14	2,13	2,33	2,41	2,4
Lysine	1,55	1,48	1,68	1,97	1,27	1,72	2,06	1,63	1,44	1,53	1,53	1,67	1,52	1,47	1,63	1,64	1,59	1,49	1,61	1,52
Methionine	0,75	0,71	0,62	0,62	0,81	0,74	0,72	0,78	0,74	0,81	0,83	0,66	1,18	0,95	0,7	0,6	0,66	0,77	0,89	0,8
Phenylalanine	1,53	1,48	1,52	1,46	1,62	1,46	1,33	1,52	1,59	1,65	1,67	1,44	1,74	1,75	1,52	1,44	1,46	1,63	1,75	1,62
Proline	1,67	1,58	1,64	1,7	1,43	1,63	1,69	1,67	1,62	1,51	1,5	1,53	1,57	1,58	1,91	1,69	1,65	1,79	1,61	1,6
Serine	1,84	1,81	1,98	2,03	1,64	1,89	1,95	1,86	1,91	1,75	1,73	1,88	1,74	1,79	1,98	2,04	1,96	1,88	1,73	1,76
Threonine	1,7	1,63	1,76	1,77	1,5	1,7	1,78	1,71	1,63	1,7	1,64	1,7	1,67	1,65	1,75	1,74	1,83	1,54	1,64	1,7
Tryptophane	0,34	0,44	0,39	0,24	0,43	0,59	0,22	0,48	0,44	0,47	0,47	0,42	0,56	0,61	0,42	0,3	0,49	0,94	0,77	0,54
Tyrosine	1,24	1,38	1,31	1,3	1,42	1,32	1,18	1,32	1,39	1,41	1,35	1,38	1,38	1,42	1,35	1,27	1,25	1,5	1,57	1,39
Valine	2,13	1,99	1,85	1,81	2,12	1,93	1,89	1,99	2,01	2,17	2,19	1,9	2,12	2,14	1,89	1,89	2,02	2,13	2,05	2,25

Table 8 Amino acid interaction percentages for mesophiles. Analysis performed at 5Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,09	1,97	1,76	1,79	1,78	1,87	1,84	1,99	1,86	1,93	1,98	1,83	1,88	1,88	1,86	1,88	1,84	1,86	1,83	2
Arginine	1,72	2,1	1,74	2,13	1,59	1,76	2,02	1,83	1,73	1,59	1,68	1,53	1,68	1,67	1,86	1,74	1,73	1,73	1,69	1,61
Asparagine	1,38	1,35	2,01	1,66	1,34	1,61	1,55	1,43	1,53	1,31	1,27	1,62	1,32	1,44	1,55	1,6	1,63	1,47	1,49	1,26
Aspartic acid	1,8	2,03	2,02	2,02	1,49	1,9	1,81	1,73	1,94	1,48	1,57	2,11	1,56	1,59	1,75	2,01	1,97	1,68	1,65	1,55
Cystein	0,2	0,03	0,21	0,09	2,15	0,15	-0,08	0,35	0,54	0,33	0,28	0,03	0,25	0,39	0,28	0,25	0,25	0,45	0,33	0,38
Glutamine	1,23	1,31	1,39	1,43	1,13	1,8	1,33	1,21	1,3	1,14	1,24	1,32	1,18	1,21	1,33	1,36	1,33	1,35	1,22	1,12
Glutamic acid	1,69	2,11	1,8	1,59	1,42	1,85	2,07	1,58	1,82	1,57	1,65	2,14	1,64	1,63	1,67	1,82	1,75	1,59	1,64	1,56
Glycine	1,65	1,63	1,68	1,73	1,51	1,54	1,56	1,85	1,63	1,55	1,51	1,59	1,55	1,58	1,85	1,76	1,78	1,67	1,57	1,64
Histidine	0,87	0,83	0,95	1,12	1,07	0,91	0,96	1	1,54	0,84	0,93	0,74	0,97	0,95	1,06	1	1,02	1,05	0,97	0,88
Isoleucine	1,9	1,7	1,72	1,56	1,97	1,67	1,63	1,82	1,76	2,26	2,04	1,79	1,98	2	1,67	1,69	1,8	1,81	1,96	2,07
Leucine	2,44	2,32	2,12	2,03	2,37	2,25	2,17	2,35	2,24	2,51	2,61	2,23	2,41	2,41	2,18	2,19	2,19	2,32	2,39	2,47
Lysine	1,43	1,26	1,7	1,93	1,27	1,59	1,94	1,52	1,43	1,48	1,41	1,79	1,43	1,44	1,44	1,59	1,54	1,45	1,56	1,41
Methionine	0,77	0,64	0,61	0,53	0,73	0,68	0,6	0,79	0,71	0,88	0,82	0,64	1,52	0,89	0,71	0,63	0,68	0,9	0,84	0,83
Phenylalanine	1,62	1,5	1,52	1,38	1,81	1,49	1,38	1,66	1,58	1,74	1,74	1,52	1,79	1,9	1,66	1,51	1,53	1,77	1,75	1,7
Proline	1,4	1,37	1,33	1,4	1,22	1,39	1,46	1,42	1,34	1,19	1,26	1,3	1,29	1,37	1,65	1,43	1,41	1,56	1,4	1,29
Serine	1,7	1,64	1,83	1,91	1,6	1,83	1,87	1,64	1,76	1,59	1,56	1,73	1,58	1,63	1,75	2	1,82	1,7	1,68	1,6
Threonine	1,69	1,59	1,76	1,82	1,6	1,82	1,84	1,69	1,7	1,68	1,64	1,66	1,65	1,66	1,68	1,75	1,89	1,61	1,69	1,69
Tryptophane	0,64	0,77	0,59	0,51	0,83	0,73	0,49	0,79	0,77	0,65	0,7	0,56	0,84	0,76	1,02	0,62	0,59	1,11	0,78	0,65
Tyrosine	1,43	1,5	1,53	1,5	1,5	1,48	1,37	1,56	1,57	1,52	1,52	1,6	1,61	1,66	1,75	1,45	1,42	1,61	1,76	1,47
Valine	2,07	1,89	1,81	1,77	2,08	1,84	1,8	2,02	1,9	2,22	2,16	1,86	2,11	2,09	1,87	1,85	1,98	2,04	2,03	2,35

Table 9 Amino acid interaction percentages for mesophiles. Analysis performed at 7Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,41	2,19	2,01	2,03	2,06	2,09	2,07	2,21	2,12	2,25	2,29	2,04	2,18	2,16	2,15	2,11	2,12	2,18	2,13	2,29
Arginine	1,5	1,57	1,43	1,81	1,34	1,58	1,89	1,59	1,5	1,43	1,5	1,35	1,45	1,43	1,56	1,47	1,42	1,53	1,5	1,45
Asparagine	1,28	1,37	1,77	1,62	1,19	1,56	1,49	1,39	1,43	1,27	1,21	1,6	1,25	1,35	1,51	1,55	1,56	1,34	1,4	1,21
Aspartic acid	1,74	2,03	2	1,84	1,35	1,86	1,8	1,67	1,85	1,42	1,47	2,11	1,48	1,46	1,74	1,98	1,9	1,57	1,51	1,48
Cystein	0,29	0,22	0,36	0,21	2,24	0,21	0,11	0,4	0,67	0,44	0,38	0,13	0,47	0,47	0,36	0,41	0,34	0,51	0,46	0,44
Glutamine	1,09	1,21	1,29	1,39	0,88	1,45	1,3	1,12	1,13	1,02	1,07	1,22	1,03	1,04	1,22	1,21	1,19	1,21	1,05	0,99
Glutamic acid	1,65	2,05	1,81	1,57	1,28	1,84	1,78	1,59	1,72	1,53	1,57	2,15	1,52	1,53	1,64	1,75	1,71	1,56	1,52	1,52
Glycine	1,87	1,93	1,96	1,97	1,84	1,83	1,82	2,08	1,9	1,85	1,82	1,81	1,92	1,9	2,11	2,02	2,04	1,96	1,88	1,89
Histidine	0,73	0,77	0,85	1,06	0,94	0,86	0,86	0,87	1,21	0,75	0,77	0,67	0,86	0,84	0,96	0,9	0,89	0,94	0,84	0,77
Isoleucine	1,91	1,72	1,75	1,62	1,98	1,72	1,67	1,82	1,81	2,12	2	1,78	2,01	2	1,63	1,73	1,82	1,81	1,95	2
Leucine	2,32	2,28	2,08	2,06	2,29	2,23	2,14	2,29	2,26	2,41	2,51	2,18	2,35	2,41	2,18	2,11	2,15	2,27	2,37	2,39
Lysine	1,39	1,28	1,6	1,89	1,17	1,6	1,99	1,51	1,31	1,48	1,4	1,63	1,41	1,41	1,47	1,54	1,48	1,34	1,53	1,43
Methionine	0,62	0,57	0,59	0,61	0,78	0,6	0,59	0,7	0,73	0,77	0,75	0,57	1,19	0,85	0,63	0,57	0,58	0,81	0,83	0,75
Phenylalanine	1,4	1,39	1,45	1,35	1,67	1,39	1,27	1,48	1,52	1,56	1,58	1,38	1,62	1,69	1,5	1,39	1,4	1,62	1,63	1,54
Proline	1,66	1,61	1,58	1,7	1,47	1,62	1,69	1,64	1,62	1,48	1,5	1,52	1,6	1,62	1,84	1,66	1,64	1,8	1,65	1,56
Serine	1,73	1,74	1,92	1,95	1,67	1,84	1,87	1,78	1,82	1,68	1,63	1,79	1,66	1,7	1,87	1,98	1,87	1,78	1,71	1,67
Threonine	1,72	1,64	1,79	1,81	1,57	1,78	1,8	1,75	1,69	1,73	1,68	1,68	1,69	1,69	1,79	1,8	1,85	1,63	1,68	1,74
Tryptophane	0,26	0,48	0,37	0,18	0,52	0,51	0,21	0,42	0,51	0,37	0,35	0,34	0,52	0,53	0,43	0,27	0,32	0,82	0,6	0,39
Tyrosine	1,17	1,33	1,36	1,21	1,4	1,34	1,15	1,3	1,39	1,38	1,35	1,41	1,42	1,43	1,33	1,23	1,24	1,43	1,55	1,35
Valine	2,22	2	1,88	1,87	2,18	1,96	1,93	2,06	2,02	2,26	2,24	1,93	2,2	2,17	1,93	1,95	2,06	2,1	2,1	2,32

Table 10 Amino acid interaction percentages for thermophiles. Analysis performed at 5Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,04	1,63	1,7	1,61	1,67	1,77	1,65	1,93	1,69	1,98	1,97	1,68	1,96	1,88	1,62	1,74	1,84	1,68	1,85	1,82
Arginine	1,64	1,99	1,86	2,12	1,39	1,8	2,06	1,77	1,7	1,66	1,69	1,58	1,98	1,54	1,77	1,81	1,83	2,01	1,77	1,58
Asparagine	1,36	1,37	2,03	1,73	1,3	1,62	1,45	1,36	1,27	1,24	1,23	1,45	1,41	1,4	1,56	1,6	1,39	1,67	1,35	1,12
Aspartic acid	1,74	2,04	1,99	1,9	1,49	1,71	1,71	1,57	1,66	1,38	1,42	2,12	1,61	1,5	1,61	1,94	2,1	1,49	1,56	1,61
Cystein	-0,18	-0,27	-0,06	-0,24	2,32	-0,33	-0,5	-0,22	-0,34	-0,54	-0,62	-0,85	-0,13	-0,44	-0,47	-0,08	0,01	-1,06	-0,15	0,12
Glutamine	0,98	1,09	1,03	1,1	0,89	1,63	0,97	0,95	1,24	0,77	0,81	0,84	1,18	1,01	1	1,18	1,1	1,06	1,27	0,87
Glutamic acid	1,83	2,32	2,12	1,81	2,06	2,18	2,18	1,71	2,31	1,83	1,95	2,45	1,84	1,77	1,95	2,04	1,77	1,79	1,68	1,67
Glycine	1,55	1,43	1,77	1,57	1	1,63	1,52	1,69	1,55	1,58	1,5	1,51	1,55	1,7	1,77	1,8	1,89	1,93	1,53	1,58
Histidine	0,49	0,72	0,67	0,83	0,77	1,01	0,63	0,68	0,96	0,58	0,61	0,69	0,64	0,75	0,92	0,75	0,84	0,55	0,83	0,66
Isoleucine	2,2	1,86	1,82	1,79	1,96	1,67	1,81	2,02	2,08	2,35	2,26	1,94	2,09	2,15	1,96	1,95	1,96	1,98	2,1	2,31
Leucine	2,43	2,26	1,86	2,03	2,29	2,16	2,08	2,49	2,19	2,54	2,53	2,31	2,35	2,49	2,21	2,08	2	2,28	2,35	2,43
Lysine	1,61	1,49	1,85	2,1	1,7	1,72	2,34	1,78	1,82	1,72	1,72	1,95	1,67	1,85	1,74	1,79	1,89	1,93	1,94	1,78
Methionine	1,03	0,66	0,68	0,52	0,77	0,39	0,61	0,7	0,9	0,77	0,61	0,49	1,17	0,87	0,72	0,52	0,67	1,06	0,89	0,71
Phenylalanine	1,74	1,56	1,44	1,52	1,79	1,4	1,55	1,74	1,52	1,62	1,7	1,53	1,54	1,79	1,87	1,67	1,45	1,53	1,79	1,6
Proline	1,32	1,44	1,2	1,3	1,08	1,43	1,46	1,24	1,33	1,15	1,13	1,21	0,97	1,31	1,46	1,41	1,1	1,38	1,43	1,17
Serine	1,37	1,67	1,4	1,75	1,37	1,51	1,57	1,45	1,3	1,22	1,33	1,42	1,38	1,35	1,48	1,65	1,66	1,33	1,3	1,47
Threonine	1,57	1,41	1,78	1,85	1,39	1,82	1,65	1,59	1,85	1,56	1,59	1,44	1,36	1,42	1,56	1,6	1,82	1,51	1,57	1,66
Tryptophane	0,52	0,77	0,54	0,55	0,7	0,93	0,64	0,75	0,88	0,36	0,63	0,4	0,77	0,89	0,93	0,6	0,37	1,24	0,76	0,74
Tyrosine	1,56	1,56	1,71	1,58	1,43	1,72	1,52	1,75	1,61	1,63	1,56	1,64	1,8	1,59	1,85	1,53	1,7	1,77	1,65	1,44
Valine	2,17	2,09	2,02	1,94	2,24	1,94	1,89	2,13	1,99	2,36	2,32	2,03	2,21	2,2	2,05	2,03	2,03	1,87	2,13	2,46

Table 11 Amino acid interaction percentages for thermophiles. Analysis performed at 7Å.

Table is read both in rows and columns. When read in columns, e.g. Alanine column, the numbers describe the logarithmic interaction percentage of alanines with alanine, then arginine etc. When read in lines, e.g. Alanine line, the numbers describe the logarithmic interaction percentage of each amino acid with alanine alone. high interaction percentages are values greater than 1.86, noted with dark green. medium interaction percentages are values greater that 1.55, noted with dark yellow.

	Alanine	Arginine	Asparagine	Aspartic acid	Cystein	Glutamine	Glutamic acid	Glycine	Histidine	Isoleucine	Leucine	Lysine	Methionine	Phenylalanine	Proline	Serine	Threonine	Tryptophane	Tyrosine	Valine
Alanine	2,29	2,04	1,98	1,86	1,88	2,09	1,85	2,14	2,2	2,28	2,37	1,91	2,22	2,1	1,96	2,09	2,1	1,95	2,2	2,25
Arginine	1,48	1,52	1,35	1,83	1,52	1,66	1,94	1,64	1,53	1,51	1,46	1,38	1,71	1,52	1,5	1,64	1,6	1,76	1,67	1,44
Asparagine	1,06	1,33	1,77	1,57	1,27	1,47	1,44	1,28	1,06	1,04	1,03	1,31	1,14	1,37	1,52	1,48	1,27	1,45	1,2	1,02
Aspartic acid	1,63	2,01	2	1,8	1,43	1,78	1,74	1,52	1,59	1,41	1,28	2,09	1,53	1,36	1,62	1,91	1,93	1,55	1,42	1,46
Cystein	-0,18	-0,48	0,09	0	2,15	-0,37	-0,5	-0,02	0,35	-0,45	-0,5	-0,4	-0,38	-0,28	-0,12	-0,04	-0,37	-0,07	-0,58	-0,11
Glutamine	0,84	1,06	0,9	1,02	0,02	1,4	1,01	0,85	1,04	0,47	0,64	0,84	1,08	0,57	0,73	0,97	0,61	0,84	0,89	0,62
Glutamic acid	1,71	2,34	2,17	1,85	1,47	2,06	2,19	1,74	2,02	1,7	1,75	2,53	1,82	1,69	1,98	2,05	1,78	1,73	1,63	1,7
Glycine	1,79	1,82	2,05	1,9	1,84	1,78	1,71	1,97	1,94	1,94	1,72	1,8	1,78	1,86	1,94	1,95	2,15	2,16	1,9	1,96
Histidine	0,29	0,62	0,82	0,79	0,9	0,92	0,5	0,55	1,03	0,28	0,55	0,45	0,58	0,68	0,82	0,56	0,34	0,71	0,95	0,49
Isoleucine	2,33	1,92	1,84	1,79	1,75	1,62	1,8	1,94	2,02	2,26	2,23	1,94	2,21	2,2	1,9	1,83	1,99	1,91	2,13	2,22
Leucine	2,36	2,2	1,89	2,16	2,17	2,19	2	2,43	2,2	2,43	2,46	2,25	2,25	2,46	2,15	2,07	1,97	2,24	2,44	2,3
Lysine	1,45	1,53	1,79	1,98	1,75	1,84	2,45	1,8	1,48	1,79	1,74	1,87	1,74	1,7	1,66	1,7	1,86	1,87	1,92	1,69
Methionine	0,9	0,5	0,64	0,59	0,71	0,57	0,67	0,75	0,75	0,9	0,66	0,51	1,11	0,99	0,62	0,49	0,71	1,03	0,68	0,8
Phenylalanine	1,6	1,31	1,37	1,49	1,6	1,45	1,25	1,46	1,41	1,44	1,55	1,47	1,55	1,73	1,62	1,53	1,37	1,73	1,59	1,59
Proline	1,46	1,54	1,62	1,71	1,51	1,54	1,61	1,58	1,65	1,38	1,51	1,38	1,45	1,66	1,84	1,66	1,48	1,59	1,67	1,56
Serine	1,45	1,69	1,49	1,74	1,45	1,49	1,6	1,57	1,47	1,29	1,36	1,53	1,35	1,46	1,66	1,76	1,73	1,18	1,45	1,47
Threonine	1,59	1,64	1,8	1,66	1,32	1,6	1,56	1,6	1,7	1,59	1,58	1,4	1,44	1,43	1,7	1,89	1,87	1,36	1,44	1,59
Tryptophane	0,08	0,44	0,43	0,41	0,79	0,7	0,17	0,4	0,58	-0,21	0,27	0,05	0,83	0,52	0,27	0,14	0,19	0,6	0,25	0,36
Tyrosine	1,34	1,36	1,48	1,35	1,69	1,68	1,31	1,42	1,44	1,48	1,44	1,46	1,47	1,36	1,55	1,13	1,53	1,66	1,3	1,28
Valine	2,37	2,09	1,96	2,01	2,4	2,03	1,98	2,21	2,18	2,43	2,39	2	2,19	2,21	2,18	2,04	2,18	2,13	2,15	2,43