Thesis for the degree
Master of Science

חבור לשם קבלת התואר
מוסמך למדעים

By
Ilana Lavie

מאת
אילנה לביא

Identification of protein residues and amino-acid properties that determine binding specificity of G protein-coupled receptors

זיהוי שיירי חלבון ותכונות של חומצות-אמינו שקובעות ספציפיות קישור של G protein-coupled receptors לליגנדים שונים

Advisors:
Dr. Yitzhak Pilpel[1]
Prof. Ronen Basri[2]

1. Department of Molecular Genetics.
2. Department of Computer Science & Applied Mathematics

מנחים:
ד"ר יצחק פלפל[1]
פרופ' רונן בצרי[2]

1. המחלקה לגנטיקה מולקולרית
2. המחלקה למדעי המחשב ולמתימטיקה שימושית

April, 2004

אייר, תשס"ד

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

מוגש למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

Acknowledgements

I'd like to use these pages to express my gratitude to my advisors, Dr. Tzahi Pilpel and Prof. Ronen Basri. Tzahi for his endless flow of new ideas, some of which are the basis for this work, and enthusiasm that clings to anyone around him and that has given me a passion for biology. Ronen for his wisdom, insights and immense knowledge. Working with Ronen has provided me with valuable tools for this work and hopefully, for works to come.

Both Tzahi and Ronen were a pleasure to work with, available for any question or request and always creating a great working atmosphere.

But mostly I would like to thank Tzahi and Ronen for the trust they expressed in me by being willing to guide me in this work against all objective reasoning, without any justification but their kindness and faith. I ought them my love for research and the opportunity to be involved in it.

Abstract

G protein-coupled receptors (GPCRs) are seven transmembrane (TM) proteins that represent the largest family of signal-transducing proteins, and are one of the most important target classes for drug discovery. Various high throughput assays are currently used to screen large compound libraries for agonists and antagonists towards GPCRs with particular medicinal significance. Yet, relatively little is known about the ligand binding site and ligand specificity of many GPCRs.

We present here three novel algorithms that combine compound binding data from high throughput screens, sequence data, measurements of amino-acid properties and optionally also ligand properties for the identification of amino acid positions, amino-acid properties and ligand properties that determine specificity of GPCRs towards ligand targets.

The current analysis covers 26 GPCRs whose ligand binding activity against more than 1000 compounds was measured. We define as ligand specificity determining position-property combinations, amino acid positions and properties such that the values of a property for the amino-acids in a position are similar among receptors with similar ligand binding profiles, and variable among receptors that display different activity profiles. We then use homology-based, and ab-initio structure prediction tools to select for residues that form a three-dimensional pocket in the receptor interior.

Among the highest scoring residues are well known GPCR ligand binding sites in addition to newly proposed specificity determining residues. Our description of a universal ligand specificity-determining pocket in this diverse family should allow better understanding of sequence-structure-function relationships in these proteins and facilitate assignment of ligands to orphan receptors.

Table of contents

Chapter 1
Introduction

G protein-coupled receptors (GPCRs) are seven transmembrane (TM) proteins that represent the largest family of signal-transducing proteins [1]. About 5% of the genes of the nematode C. elegans and 1% of the mammalian genome encode for GPCRs, and over 1000 GPCRs in human direct responses to an enormous diversity of signal molecules, including hormones, neurotransmitters, light photons, taste ligands and odorants. The ligand molecules have wide variety of structures, including biogenic amines, amino acids, peptides, lipids, nucleotides, and large polypeptides. [2]

GPCRs transduce information from extracellular stimuli into the cell interior through coupling to heterotrimeric G proteins, that consist of α, β and γ subunits. The G proteins are attached to the cytoplasmic face of the plasma membrane, where they serve as relay molecules, functionally coupling the receptors to either plasma-membrane-bound enzymes or ion channels. In an inactive state, both α and γ subunits have covalently attached lipid molecules that bind them to the plasma membrane, and the α subunit has GDP bound. As the agonist binds to the receptor, the α subunit exchanges the GDP with GTP. The exchange leads to a release of the $\beta\gamma$ subunits complex and subsequently to the activation of both components – the α subunit adopts a new shape that allows it to interact with its target proteins, and the surface of the $\beta\gamma$ complex, previously masked by the α subunit, interacts with a second set of target proteins. After a G-protein α subunit activates its target protein, it shuts itself off by hydrolyzing its bound GTP to GDP. This inactivates the α subunit, which dissociates from the target protein and re-associates with the $\beta\gamma$ complex to re-form an inactive G protein. Since the α subunit is a GTPase, the inactivation process occurs automatically after a maximum of several minutes. Practically, the inactivation is usually much faster due to enhancement of the GTPase activity by binding of the target protein or a specific modulator known as RGS (regulator of G protein signaling). [4]

GPCRs have been found to be dysfunctional/dysregulated in a growing number of human diseases and have been estimated to be the targets of more than 40% of all marketed drugs [2]. Cholera is an example for a disease that relates to G proteins. Cholera in caused by a toxin that alters the α subunit of a stimulatory G protein, so that it cannot become inactive. The G protein activates adenylyl cyclase and thereby increases cyclic AMP concentration. The final result of this process is a severe diarrhea that characterizes cholera. Abnormal bone development and mental retardation can be caused by a genetic deficiency in a particular stimulatory G protein (a G protein that increases cAMP concentration). Examples for drugs that target GPCRs include antidepressants such as fluoxetine (Prozac). Orphan GPCRs are viewed as potential drug targets for various diseases, including obesity, cardiovascular disease, inflammation and cancer [2]. Thus, understanding how GPCRs function at the molecular level is an important goal of biological and applied pharmaceutical research.

The 3D structure of the first GPCR, rhodopsin, has been resolved [5]. Yet in the absence of experimentally determined 3D structures of other GPCRs, especially those with only remote homology to rhodopsin, computational approaches for modeling their structure, that potentially incorporate the vast amounts of experimental data gathered for these proteins, are needed. In particular, a great challenge is the identification of functional sites of GPCRs, and delineation of the residues that determine specificity towards agonists. Such knowledge may have profound effect on two central pharmaceutical efforts, namely the design of drugs towards particular molecular targets, and the identification of potentially deleterious single nucleotide polymorphism in human populations that occur in the GPCRs' functional sites.

In this work we used data on ligand binding preferences of GPCRs in order to identify the amino acid positions that determine their specificity towards their ligands, and more specifically - the amino-acid properties that, combined with the amino-acid positions, affect binding.

Four data sets were used: measurements of ligand binding activity w.r.t 26 GPCRs and 1068 putative ligands; the values of 56 properties of the 20 amino-acids; sequence alignment of the 26 GPCRs and the values of 164 properties of each of the 1068 ligands.

For each position in the alignment and each amino-acid property, a 'position-property' vector was created – a vector of size 26, containing the values of the property w.r.t the amino-acids in the position.

Our first working hypothesis is that a position-property vector, if created by a ligand specificity-determining position and amino-acid property, would display similarity among receptors with similar ligand binding preferences, and difference among receptors with different binding activities.

Based on this hypothesis the 'Correlated distances' method was developed. The 'Correlated distances' procedure defines a measure of mutual similarities between each position-property combination in a multiple alignment, and a corresponding measure of the similarities between the ligand binding preferences of the GPCRs. We then seek for amino acid position-property combinations in which the pattern of amino acid property similarity across all pairs of proteins correlates with the pattern of similarity of the ligand binding preferences. A statistical model was constructed that identifies amino acid positions with significantly high correlation with the ligand binding data. Such positions can be regarded as candidate ligand specificity-determining residues.

Additional hypothesis is that binding specificity has a hierarchical nature. Amino-acid position and property combination can induce a basic classification of GPCRs into different binding preferences classes (for example the aminergic binding-site combined with the amino-acid property 'Charge' induces a classification of GPCRs into an aminergic class vs. non-aminergic class), while differences between binding preferences within a class can be induced by a different position-property combination (for example, serotonin class vs. histamine class). In fact, the hierarchical nature can result from Boolean properties with different induction levels.

Based on the two hypotheses a second method - 'Hierarchical clustering' method - was developed. The 'Hierarchical clustering' method assembles the GPCRs into a binary tree according to their binding specificities. Every split in the tree divides a subgroup of the GPCRs into two groups, such that each group contains GPCRs with similar

binding specificities. For every split in the tree we seek for amino acid position-property combinations, such that the property's values (w.r.t the amino-acids in the position) are relatively similar for GPCRs included in one group and different for GPCRs in different sides of the split. Every such position-property combination is a candidate ligand specificity-determining residue for the GPCRs included in the split it refers to.

A third hypothesis our work is based on, is that the effect of a position-property combination on binding specificity is mediated by some ligand property. For example, if the charge of the amino-acid in the expected position affects ligand binding activity, then binding would be affected by the charge of the ligand as well.

The third method suggested in this work, 'Correlated properties' method, relies on this hypothesis and on the fourth set of data, containing 164 ligand properties assigned to the same set of ligands used in the binding specificity data set.

The ligand properties data was analyzed to create a new binding specificity data set, containing binding preferences of the GPCRs with respect to the ligand properties, and not the ligands themselves as in the original data. The measure of binding preference of a GPCR to a ligand property was defined as the correlation between the GPCRs' ligand binding activity and ligand property values, both w.r.t the same set of ligands.

The 'Correlated properties' method seeks for triplets of position, amino-acid-property and ligand-property, such that amino-acid property values of the amino-acids in the position, taken from all GPCRs, are highly correlated with the binding preferences of the GPCRs to the ligand property. We thus look for amino acid positions that w.r.t particular properties show enhanced similarity in receptors that bind ligands of similar chemical nature.

Another hidden assumption is that ligand-properties matching the found amino-acid-properties exist in the data set..

The last method suggested in this work, 'Two-way clustering', is not a stand-alone method, but an improvement that can be used with any of the other methods. It is based on an expanded version of the hypothesis that suggests that binding specificity has a hierarchical nature. The original hypothesis states that position-property combinations can affect binding-specificity of one subgroup of GPCRs while another subgroup would not be affected by them. The last method offers an alternative procedure that finds such subgroups, and also applies this assumption to ligands as well – a position-property combination can affect binding to one subgroup of ligands, while another subgroup of ligands would not be affected by it.

'Two-way clustering' uses existing gene expression clustering algorithms to find such subgroups of GPCRs and ligands. Activation of the two last methods on a cluster, that contains only GPCRs and ligands affected by a specific position-property combination, increases the probability of finding that combination and also allows finding combinations that affect binding of small groups of GPCRs and/or ligands.

'Two-way clustering' can be combined with the first method, 'Hierarchical clustering', as well, replacing the original hierarchical clustering, which induces rigid clustering, by a more flexible procedure.

After a list of candidate positions and properties is created by one or few of the suggested methods, further analysis was performed to retrieve various properties for

each of the candidate positions. Structural properties were retrieved and are presented for the positions found by each method. These properties include: is it included in a transmembranal helix; which helix; does it face the membrane or the interior of the receptor bundle; is it located closer to the extracellular milieu or the intracellular milieu.

Chapter 2
Methods

2.1 The input data

The input data includes four sets of data. Except for the general amino-acid properties data set (see chapter 2.1.2) all the data sets relate to a set of 26 GPCRs and 1068 ligands. The GPCRs are varied and include receptors to bio-amines, neuropeptides and other substances.

The ligand binding activity data set (see chapter 2.1.1) and ligand properties data set (see chapter 2.1.4) were generated in Pfizer inc. and delivered to our lab through a research agreement. The actual molecular formula of the assayed ligands and precise mode of binding measurements were not disclosed.

2.1.1 Ligand binding activity of 26 GPCRs and 1068 ligands

Measurements of ligand binding activity of 26 G-protein coupled receptors with 1068 ligands are shown in Figure 1. The values are surrogates to the actual affinities between the GPCRs and ligands (details of the 26 GPCRs are available in http://longitude.weizmann.ac.il/GPCRs/GPCRs.html).

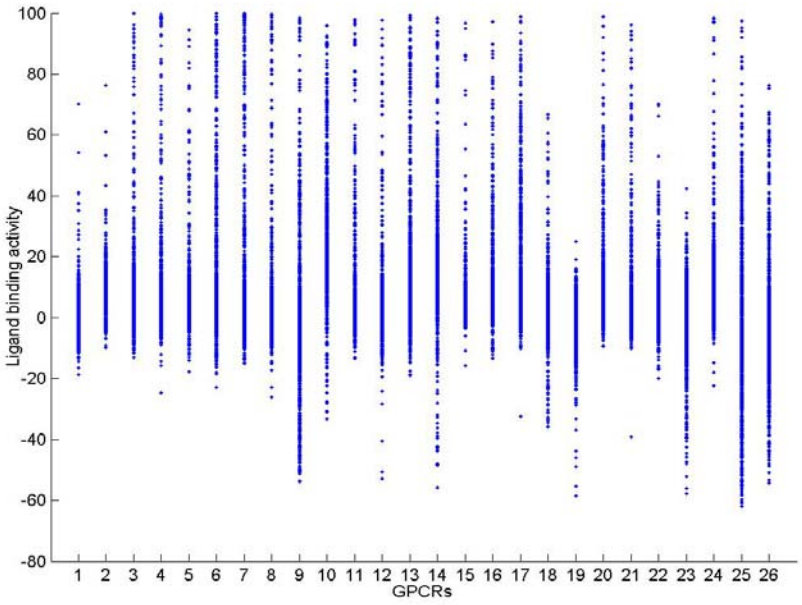


Figure 1: For each G-coupled protein receptor, each ligand measurement is displayed as a dot. The values are measurements of ligand binding activity for 26 GPCRs and 1068 ligands. The original data set includes outliers – measurements with higher variance than the majority of the data. The outliers are not presented in the figure.

The original data set includes outliers (not shown in Figure 1). Outliers can arise from errors, such as human errors or errors that originate in the nature of the experiment, or can represent the true nature of the data. In the first case, outliers should be discarded, though in the second - they should be left untouched. Since the nature of the experiment is not known, there is no reason to prefer one approach over the other, and therefore both approaches were tried.

It is a common practice to normalize biological data, since usually measurements are acquired at different conditions, and therefore may exhibit different data ranges and non-uniform noise. Therefore the data was normalized by first deducting from each value the mean (of all ligands w.r.t one GPCR) and then dividing it by the standard deviation.

2.1.2 Amino-acid properties

Values of 56 amino-acid properties w.r.t the 20 amino-acids (the list of amino-acid properties with their values w.r.t the amino-acids can be found in: http://us.expasy.org/cgi-bin/protscale.pl?143S_HUMAN). An additional amino-acid property, 'charge', was added as the 56^{th} amino-acid property. 'Charge' is equal to 0 for all amino-acids except for D and E with a value of -1, R and K with value of 1 and H with a value of 0.5). The original set of properties was assumed to include related properties, such as polarity and hydrophobicity, which are known to be highly inter-dependent. Since the various algorithms used throughout the work score each property separately by using various statistical methods, and since most statistical methods require no dependencies within the data, it is desired to remove or lessen the inter-dependency as much as possible. This was achieved by clustering the amino-acid properties into groups of highly-correlated properties and creating a new set that includes only one set of data for each cluster. The new set was created by normalizing each amino-acid property's values, and then calculating the mean vector for each cluster. The result is a set of 19 amino-acid properties' clusters, shown in Figure 2. Both the original set and the new set were used throughout the work..

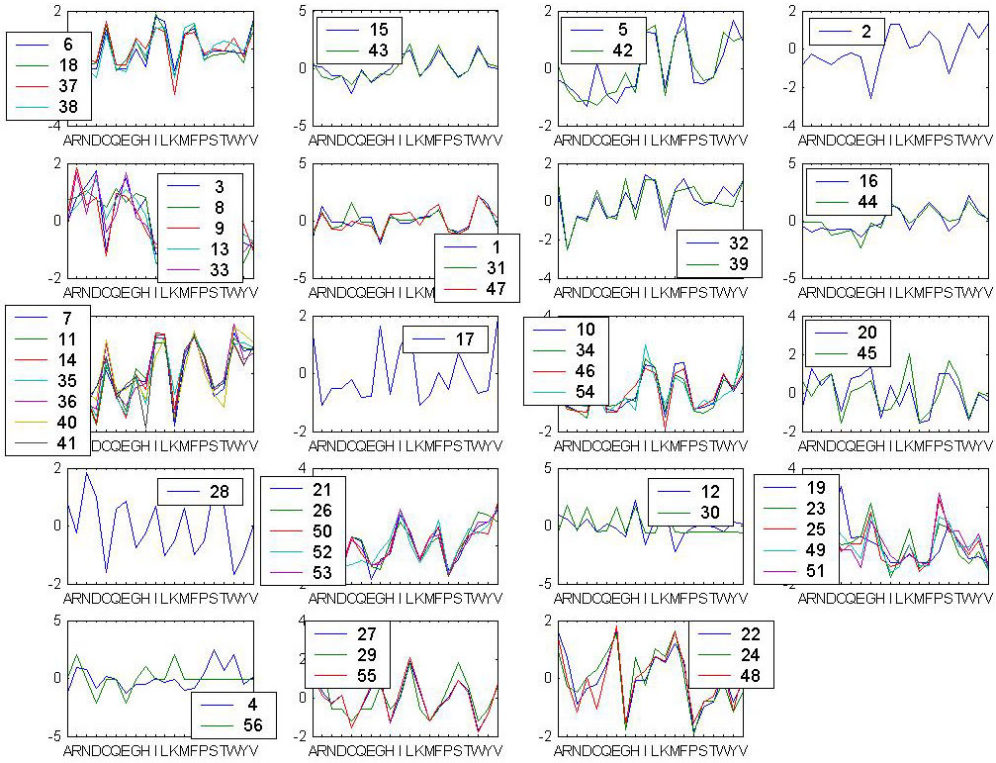


Figure 2: Clustering of amino-acid properties. To eliminate inter-dependencies between the given 56 amino-acid properties, they were clustered into 19 clusters. The distances between the amino-acid properties used for the creation of the clusters are based on the Pearson correlation coefficients between the properties' values w.r.t to the 20 amino-acids. Each plot displays the normalized values of the properties in one cluster. The legends display the indices of the amino-acid properties according to the following numbering: 1. Molecular weight; 2. Bulkiness; 3. Polarity; 4. Recognition factors; 5. Optimized matching hydrophobicity; 6. Hydropathicity; 7. Hydrophobicity (delta G1/2 cal); 8. Hydrophobicity (free energy of transfer to surface); 9. Hydrophobicity scale based on free energy of transfer (kcal/mole); 10. Hydrophobicity scale (contact energy derived from 3D data; 11. Hydrophobicity scale (pi-r); 12. Antigenicity value X 10; 13. Hydrophilicity scale derived from HPLC peptide retention times; 14. Hydrophobicity indices at ph 7.5 determined by HPLC; 15. Retention coefficient in HFBA; 16. Retention coefficient in HPLC, pH 2.1; 17. Molar fraction (%) of 2001 buried residues; 18. Proportion of residues 95 percent buried (in 12 proteins); 19. Atomic weight ratio of hetero elements in end group to C in side chain; 20. Average flexibility; 21. Conformational parameter for beta-sheet; 22. Conformational parameter for alpha helix; 23. Conformational parameter for beta-turn; 24. Normalized frequency for alpha helix; 25. Normalized frequency for beta-turn; 26. Conformational preference for antiparallel beta strand; 27. Overall amino acid composition; 28. Relative mutability of amino acids (Ala=100); 29. Number of codon(s); 30. Polarity; 31. Refractivity; 32. Normalized consensus hydrophobicity scale; 33. Hydrophilicity; 34. Average surrounding hydrophobicity; 35. Hydrophobicity of physiological L-alpha amino acids; 36. Hydrophobicity scale (pi-r); 37. Free energy of transfer from inside to outside of a globular protein; 38. Membrane buried helix parameter; 39. Hydration potential; 40. Hydrophobic constants derived from HPLC peptide retention times; 41. Hydrophobicity indices at ph 3.4 determined by HPLC; 42. Mobilities of amino acids on chromatography paper (RF); 43. Retention coefficient in TFA; 44. Retention coefficient in HPLC, pH 7.4; 45. Molar fraction (%) of 3220 accessible residues; 46. Mean fractional area loss (f) [average area buried/standard state area]; 47. Average area buried on transfer from standard state to folded protein; 48. Conformational parameter for alpha helix (computed from 29 proteins); 49. Conformational parameter for beta-turn (computed from 29 proteins); 50. Conformational parameter for beta-sheet; 51. Conformational parameter for coil; 52. Normalized frequency for beta-sheet; 53. Conformational preference for total beta strand (antiparallel+parallel); 54. Conformational preference for parallel beta strand; 55. Amino acid composition (%) in the Swiss-Prot Protein Sequence data bank; 56. Charge.

A new matrix that represents the amino-acid properties was created based on this clustering. It includes 19 elements (instead of 56), where each element is the mean of the normalized values of the amino-acid properties included in one cluster.

2.1.3 Sequence alignment of the 26 GPCRs

Sequence alignment of the 26 GPCRs. Since the purpose of this work is to find positions that can explain variability within binding-profiles of the GPCRs, positions that are not expected to affect binding were removed. Consequently two sets of positions were removed; (a) Positions that include the same amino-acid for all GPCRs. (b) Positions that are not part of the seven transmembranal helices. This step is not needed for the correctness of the various processes but shortens their running-time. (available in http://longitude.weizmann.ac.il/GPCRs/alignment.html)

2.1.4 Ligands properties

Values of 164 ligands' properties w.r.t the 1068 ligands. This data set allows an alternative approach to the definition of 'binding specificity' – instead of analyzing the preference of different GPCRs to different ligands, it is possible to analyze the preference of different GPCRs to ligands' properties, for example: to what extent does the GPCR 'A1(h)' prefer binding to hydrophobic ligands over hydrophilic ones.

The preference of each GPCR w.r.t each ligand property was defined as the correlation coefficient between the GPCR's ligand binding activity values (w.r.t binding of 1068 ligands) and the property's values w.r.t the ligands, calculated in the following way:

Denote by A the ligand binding activity matrix (details in chapter 2.1.1). A(x,y) is the measured ligand binding activity of GPCR x following the binding of ligand y.

Denote by L the Ligands' properties matrix. L(y,z) is the measured value of ligand-property z w.r.t ligand y.

Denote by C the new correlation coefficients data set.

$$C(x,y) = corrcoef\left(A(x,1:1068), L(1:1068,y)\right)$$

$$corrcoef\left(\bar{v},\bar{u}\right) = \frac{\mathrm{cov}(\bar{v},\bar{u})}{\sqrt{\mathrm{var}(\bar{v})\cdot\mathrm{var}(\bar{u})}}$$

The new correlation coefficients data set is shown in Figure 3. The p-values for each correlation were calculated and analyzed by the False Discovery Rate method [6] and significant correlations were colored red. It can be seen that the number of significant results is much higher than expected by random – ~20% of the correlations are significant (852 out of 4262). The same analysis, performed after randomly shuffling the ligand-property's vectors, yielded only 6 significant results out of 4262.

For example, the dot marked by an arrow is the correlation coefficient between the measured activity of GPCR 17 (M4(h)) following the binding of 1068 different ligands and the values of ligand-property 5 (pka(MB)) w.r.t the same set of ligands. The high correlation - 0.2340 – suggests that the binding of ligands to GPCR 17 is affected by the ligands' PKA.

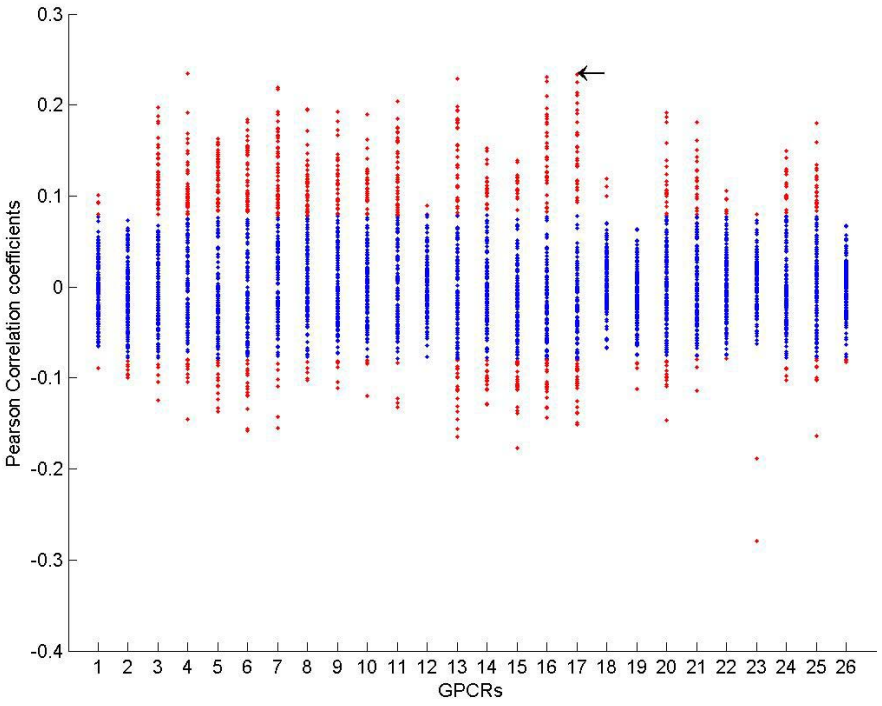


Figure 3 : For each G-coupled protein receptor and for each ligand property, the Pearson correlation coefficient between the GPCR's activity following the binding of 1068 different ligands and the property's values for the 1068 ligands was calculated. Each correlation coefficient is displayed as a dot. The p-values related to the correlation coefficients were calculated and False Discovery Rate adjustment with significance threshold of 0.05 was applied, yielding an adjusted significance threshold of 0.01 and 852 significant results (out of 4264), marked by a red color. A similar analysis on the date after a random shuffle yielded only 6 significant results. Hence, on average, the strength of the response of a GPCR to the binding of a ligand is affected by the ligand's nature as measured by ~20% of the given ligands' properties.

In addition, it was found that different GPCRs show different preferences to different ligands' properties, both referring to the identity of the properties and the number of significant properties found. For example, GPCRs A1(h)[1], A2a(h)[2], D2(h)[12], CCKB[18], Y1(h)[19], delta(h)[22], AT2(h)[23] and P2X[26] have an average of 4.4 significant ligands' properties while the other GPCRs – an average of 45.4. This result indicates that GPCRs show clear and varied preference towards specific ligands' properties.

This new data set will be used as an alternative to the original ligand binding activity matrix when referring to GPCRs' binding profiles.

2.1.5 Treatment of gaps

The sequence alignment includes numerous gaps, all located in positions which are not part of the transmembranal helices. All the methods below retrieve the amino-acids of the GPCRs in specific positions, and their values w.r.t amino-acid properties. When some of the GPCRs have gaps, they are ignored and are not used for the specific related calculations.

Nevertheless, in order to relate to the existence of gaps and to the hypothesis that it might affect binding as well, an additional property named 'Is gap' was added. The value of this property is 0 for all amino-acid and 1 for gaps.

The analysis for each position includes analysis for all original amino-acid properties, relating only to GPCRs that do not have a gap in the position, and an additional analysis for the 'Is gap' property that relates to all GPCRs.

In 'Correlated properties' method, cases may occur where for a specific position and amino-acid property a value must be set for all GPCRs including GPCRs that have a gap in the specific position. These cases have a special treatment, details in the chapter that describes the method (chapter 2.4.2).

2.2 Method 1 - Hierarchical clustering

2.2.1 Introduction

The 'Hierarchical clustering' method is based on the hierarchical nature of binding specificity. For example, mutagenesis and other experiments revealed that binding of aminergic GPCRs with all aminergic ligands involves a direct contact between the highly conserved Asp residue in the third transmembranal helix and the protonated amine of the ligand [12]. In addition, it was found that binding of β-adrenergic receptor with the antagonist iodocyanopindolol doesn't occur if the negatively charged aspartate residue at the 3rd transmembranal helix is replaced with an uncharged asparagines residue [12] . Therefore, division of a group of GPCRs into sub-groups according to whether the amino-acid in the above position is charged or not is expected to yield groups with different binding profiles. Furthermore, if we divide only the GPCRs included in one of the groups into sub-sub-groups, we might find a different position and amino-acid property that affects binding profiles within the group.

2.2.2 The method

The hierarchical clustering was created by the 'linkage' function of Matlab. The function takes as an input a list of pairwise distances and creates the hierarchy in an iterative process, where at each step the two closest objects are combined, and the pairwise distances vector is updated to include distances to the newly joined cluster instead of to the original elements. In this analysis the 'ward' option was used - the distances to the new cluster are calculated as the mean of distances to the two elements before the join minus the distance between the two joined elements, weighted by the number of elements in each cluster.

The calculation of the distance of an element (or cluster of elements) to the newly joined cluster is as follows:

Denote the number of elements in a cluster a by n_a and the distance between clusters a and b by R(a,b). If cluster x and cluster y are joined, the distance between an element/cluster z to the newly joined x-y cluster, is:

$$\sqrt{\frac{(n_x + n_z)R(x,z) + (n_y + n_z)R(y,z) - n_z R(x,y)}{n_x + n_y + n_z}}$$

The pairwise distances are Euclidean distances between the binding profiles of pairs of GPCRs. Two hierarchical clustering trees were created: (1) The binding profile of a GPCR is its normalized ligand binding activity vector. (2) The binding profile is a vector of correlation coefficients between the GPCRs' ligand binding activity and the values of 164 ligand properties (details in chapter 2.1.4). Shown in Figure 4.

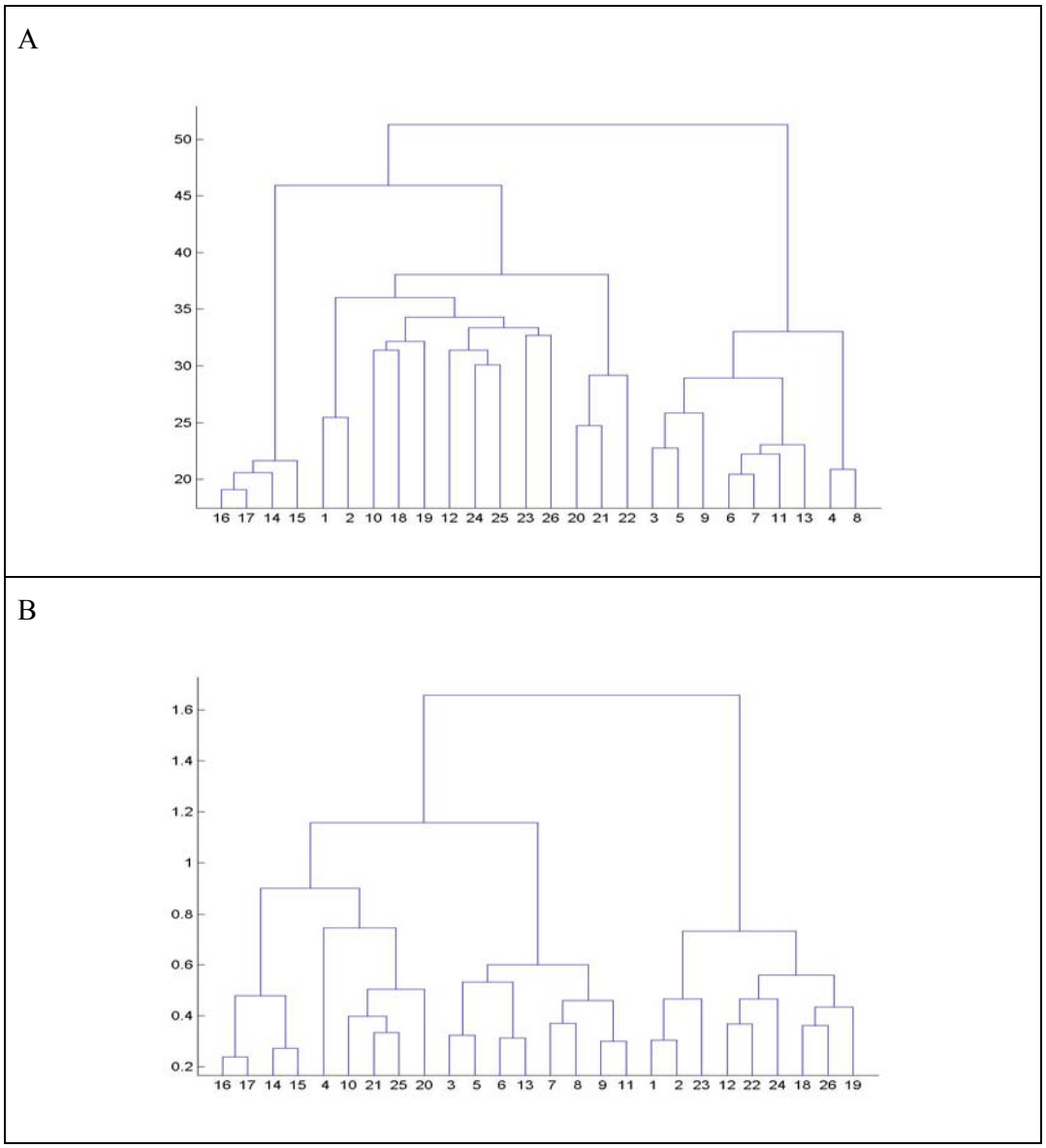


Figure 4: Hierarchical clustering of the 26 GPCRs. The distances between the GPCRs used for the creation of the dendrograms are Euclidean distances. **(A)** Each GPCR is represented by its normalized measured activity values following binding of 1068 ligands. **(B)** Each GPCR is represented by a vector of size 164, where the i'th element is the Peasron correlation between the measured activity of the GPCR following binding of 1068 ligands and the values of the i'th ligand property w.r.t to the 1068 ligands.

The objective of the algorithm is to identify, for each split in the dendrogram, a position in the alignment of the GPCRs and an amino-acid property that might be the

reason for the difference between binding profiles of GPCRs in the different sides of the split. In other words – to find a position and amino-acid property, such that the value of the property w.r.t the amino-acid in the position affects the binding profile, and causes the GPCRs in one side of the split to have different binding profiles from the ones in the other side.

For every split in the dendrogram, two groups of GPCRs are created – GPCRs that are descendants of the right side of the split, and GPCRs that are descendants of the left side of the split. For every split, the algorithm iterates through all positions and retrieves for each position the amino-acids of the GPCRs in the two groups. For each position, the algorithm iterates through all amino-acid properties and retrieves each amino-acid property's values w.r.t to the amino-acids found.

Hence, for every combination of position and amino-acid property a pair of vectors is created, holding the values of the property for the amino-acids in the position for the GPCRs in the two groups.

For every position-property combination, the two-sample Kolmogorov-Smirnov test is applied to the two vectors to determine the p-value related to the null hypothesis that the two vectors have the same continuous distribution.

After iterating over all positions and over all amino-acid properties for one split, FDR is applied to the resulting p-values with significance threshold 0.05, to adjust the threshold to the number of position-property combinations.

The results of this algorithm are lists of position-property pairs, such that each list 'explains' one split in the dendrogram and sets a condition for classifying GPCRs into one of the two groups induces by the split.

2.2.3 Discussion

Hierarchical clustering

The use of dendrograms fits the nature of the data and of the requested results, by allowing the discovery of position-property combinations that affect only partial groups of GPCRs. Nevertheless, the results are highly sensitive to the initial hierarchical clustering; different classification of numerous GPCRs can cause an extreme change in the final results. There are numerous ways of creating such dendrograms, each yielding different results.

Therefore, a bootstrap method [7] was used to measure the stability of the different splits. The calculation is done as follows:

1. Denote the original data set used for the hierarchical clustering – data set. Data set is a matrix that holds a vector of size n for each GPCR.
2. Alter the original data-set by sampling (with repetitions) n values from the original set, and using the sampled elements instead of the original ones. Re-create the dendrogram with the altered data set.
3. For each split in the original dendrogram, check whether it is present in the new dendrogram (contains the same list of GPCRs in the right and left sides).
4. Repeat steps 2-3 1000 times.

The stability of a split is estimated by the percentage of altered dendrograms that include it.

The bootstrap analysis was performed on the dendrograms, and only stable splits were analyzed by the algorithm.

Scoring the position-property combinations

The scores for the significance of each position-property combination are the p-values associated with the two-sample Kolmogorov-Smirnov test. The main reason for choosing this test over Ttest or Anova is that it is not dependent on the data set to be normally distributed, and most of the amino-acid properties are indeed not normally distributed.

The Kolmogorov-Smirnov test tests differences between two distributions by calculating the unsigned differences between the relative cumulative frequency distributions of the two samples.

Denote one distribution by X1, and a second distribution by X2.

For each potential value x, the Kolmogorov-Smirnov test compares the proportion of X1 values less than x with proportion of X2 values less than x. The test-statistic is equal to the maximum difference over all x values. Mathematically, this can be written as:

$$test - statistic = \max(|F1(x) - F2(x)|)$$

where F1(x) is the proportion of X1 values less than or equal to x and F2(x) is the proportion of X2 values less than or equal to x. The p-value is calculated according to the test-statistic and the sizes of the two distributions.

Expected critical values can be looked up in a table or approximated. Comparison between observed and expected values leads to decisions about whether the maximum difference between the two cumulative frequency distributions is significant.

2.3 Method 2 - Correlated distances

2.3.1 Introduction

The 'Correlated distances' method is based on the assumption that if binding specificity is determined by the charge property of the amino-acid in position x, then GPCRs with similarly charged amino-acid in position x have similar binding profiles, while GPCRs with differently charged amino-acid in the same position have different binding profiles.

2.3.2 The method

The algorithm iterates until no significant results are found. Each iteration is as follows:

1. For each group of GPCRs found by the previous iteration (for the first iteration use the group of all GPCRs in the data set):

1.1. Calculate pairwise Euclidean distances between the binding profiles of every pair of GPCRs. The binding profile can be defined in one of two ways: either the normalized ligand binding activity values, or the correlation between each GPCR's ligand binding activity values and the given 164 ligand properties (see chapter 2.1.4) which represents the extent to which its binding is affected by 164 ligand properties.

1.2. For each position:

1.2.1. Retrieve the amino-acids of the GPCRs in the current position.

1.2.2. For each amino-acid property:

1.2.2.1. Create a vector of the amino-acid property's values w.r.t the amino-acids vector created in step 1.2.1.

1.2.2.2. Calculate pairwise differences between the elements of the vector.

1.2.2.3. Calculate the correlation coefficient between the binding distances and the property's distances.

1.2.2.4. Calculate the p-value associated with the correlation by repeatedly shuffling the property's values vector, creating a new random differences vector and correlating it with the binding distances. The p-value is estimated as the percentage of shuffles that yielded a correlation coefficient larger or equal to the original one.

1.3. Step 1.2 creates a matrix of p-values, where each p-value is associated to one position and one amino-acid property. Apply FDR to the p-values with significance threshold 0.05.

1.4. If the position-property combination with the lowest p-value is significant (according to FDR), divide the GPCRs into two groups according to whether their values w.r.t that combination are higher or lower than the average value.

1.5. Add the two GPCRs subgroups created in step 1.4 to the list of subgroups to be analyzed by the next iteration.

2.3.3 Discussion

General

The main advantage of the 'Correlated distances' method is its robustness, which is achieved because it does not use the hierarchical clustering (or any other clustering mechanism), and thus it is insensitive to its results.

Calculation of the p-values

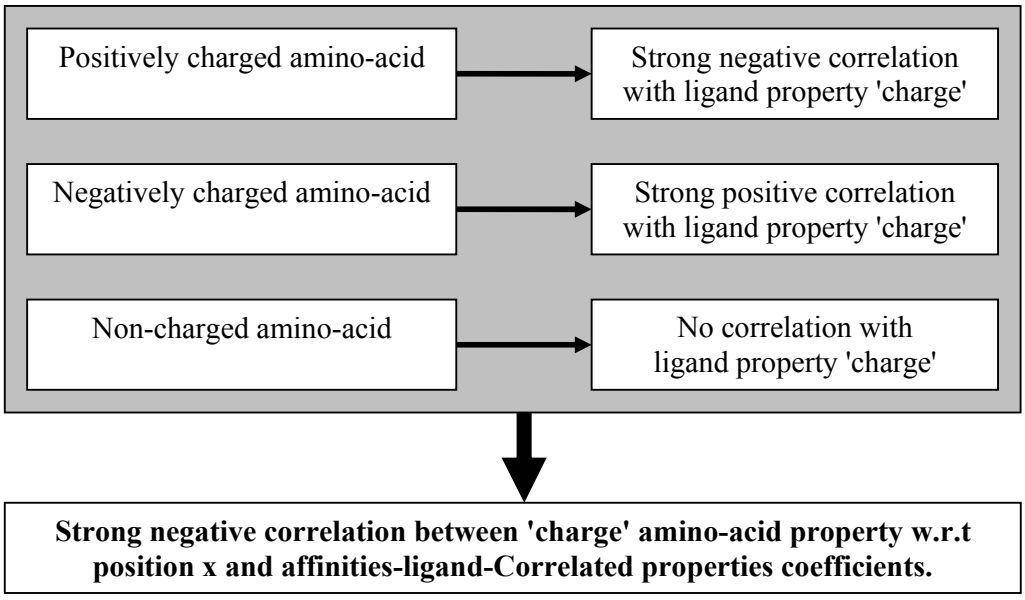
Scoring the position-property pairs is done by calculating the p-values of the correlation between pairwise distances of binding profiles and pairwise distances of the properties' values .The common method for calculating the p-value associated with a correlation coefficient transforms the correlation coefficient to create a t-statistic having N-2 degrees of freedom, where N is the number of elements in each correlated element.

This method, however, was not used due to the fact that the correlated data contains pairwise distances, and therefore incorporates inter-dependencies. Instead, a straightforward sampling of randomly shuffled data was used.

2.4 Method 3 - Correlated properties

2.4.1 Introduction

The 'Correlated properties' method searches for triplets of position, amino-acid property and ligand property that affect binding. It is based on the assumption that the position-property combination that affects binding is highly correlated with some ligand property. For example, if binding is dependent on the charge of the amino-acid in position x, it is reasonable to assume that it is also dependent on the charge of the ligand – positively charged amino-acid would bind strongly negatively charged ligands but would not bind at all positively charged ligands, and vice versa. In such a case, there should be a high (negative) correlation between the charge value of the amino-acids in position x of the GPCRs and the correlation coefficients between the ligand binding activity values of the GPCRs and the values of the ligand property 'Charge' (see chapter 3.1.4).



Therefore, the algorithm looks for position – amino-acid-property – ligand-property triplets that show high correlation.

In addition, 'Correlated properties' method tries to find secondary triplets – triplets that affect binding, but to a lesser extent than the major triplets. Such triplets would not yield high scores when calculated with the original data sets, since the major triplets mask their effect, but can yield high scores from data sets, for which the effect of the major triplets is 'cancelled'.

'Correlated properties method is an iterative procedure; in each iteration the procedure:

1. Finds triplets of position, amino-acid property and ligand-property that show significant correlation between the values of the amino-acid property w.r.t the amino-acids in the position, and the set of correlation coefficients between the measured activity of the GPCRs following binding of the different ligands and the values of the ligand-property w.r.t the ligands.
2. 'Cancels' the effect of the triplets found in step 1 by changing the original ligand binding activity data set.

2.4.2 The method

The algorithm iterates until no significant triplets are found. In each iteration, the procedure:

1. Creates a matrix of size equal to [no. of GPCRs]×[no. of ligand-properties] (26×164). Each value in the matrix is equal to the Pearson correlation coefficient between the measured activity of a GPCR following binding of 1068 ligands and the values of a ligand-property w.r.t the same 1068 ligands. Every value represents the extent to which the ligand binding activity of a GPCR is affected by a ligand property.
2. Each position, amino-acid property and ligand property triplet is scored. The score is equal to the p-value associated with the Pearson correlation between the values of the amino-acid property w.r.t the position, and the vector of correlation coefficients related to the ligand property that was created in step 1.
3. Only triples with p-value <= 0.05 FDR corrected are considered significant.
4. The list of significant triplets is analyzed to cancel interdependencies between the different amino-acid properties and between the ligand-properties.

   For every position that was found significant (included in at least one of the triplets), a list of all amino-acid properties attached to it (included in a triplet with it) is created. The pairwise correlation coefficients between the amino-acid properties (calculated upon the vectors of their values w.r.t the 20 amino-acids) are calculated, as well as the p-values associated to them. The list is clustered using the Matlab function 'cluster'. The clusters are created such that two amino-acid properties are included in the same cluster only if the p-value associated with their correlation coefficient is smaller than 0.1.

   Each amino-acid property in the cluster is scored by the mean of the p-values attached to the correlation between it and the other amino-acid properties in the cluster. This score is an estimate to how close the amino-acid property is to the center of the cluster, and therefore to how well it 'represents' it.

   For every amino-acid properties cluster found, a similar procedure is used to cluster the ligand properties attached to it.

   For every triplets cluster (for which all positions are equal, all amino-acid properties belong to the same cluster as well as the ligand properties), only one triplet is selected. The selected triplet includes the amino-acid property and ligand property that the multiplication of their scores is minimal.
5. The effect of each of the remaining triplets is canceled by the following procedure:

a. For every GPCR, the linear polynomial regression line of the curve, created by the ligand-property values (w.r.t 1068 ligands) in the x-axis and the ligand binding activity of the GPCR (to 1068 ligands) in the y-axis, is calculated.

   Denote by (pos, aaProp, ligProp) the triplet found by the procedure.

   Denote by $\bar{a}(G)$ the ligand binding activity values of GPCR 'G' w.r.t 1068 ligands.

   Denote by $\bar{l}$ the values of ligProp w.r.t the same 1068 ligands.

   Denote by 'regression line 1' the linear polynomial regression line that fits, for each GPCR 'G', $\bar{a}(G)$ to $\bar{l}$:

$$\bar{a}(G) = p_1 \times \bar{l} + p_0$$

b. The linear polynomial regression line of the curve, created by the amino-acid property (w.r.t the amino-acid in the position for all GPCRs) in the x-axis and the slopes of the regression lines created in step 5.a. in the y-axis, is calculated.

   Denote by $p_1(G)$ the slope of regression line 1 calculated for GPCR 'G', and by $\bar{p}_1$ the vector of slopes of the regression lines calculated for all GPCRs.

   Denote by c(G) the value of aaProp w.r.t the amino-acid in pos of GPCR 'G', and by $\bar{c}$ – the vector of values of aaProp w.r.t the amino-acids in pos of all GPCRs.

   Denote by 'regression line 2' the linear polynomial regression line that fits $\bar{p}_1$ (in y-axis) to $\bar{c}$ (in x-axis):

$$p_1(G) = q_1 \times c(G) + q_0$$

   If some of the GPCRs have gaps in the selected position, they are ignored for the creation of regression line. Once the regression line is set, their amino-acid property values are estimated according to their slope ($p_1(G)$ - calculated in step a) and the regression line.

   The procedure adjusts $\bar{p}_1$ and creates a new set of adjusted slopes, such that the effect of $\bar{c}$ will not be reflected. Denote the vector of adjusted slopes as $\bar{p}_2$. $\bar{p}_2$ is equal to the residuals of regression line 2:

$$p_2(G) = p_1(G) - (q_1 \times c(G) + q_0)$$

c. The original ligand binding activity values are adjusted such that the effect of the triplet (pos, aaProp, ligProp) is neutralized.

   Denote by $\hat{a}(G)$ the adjusted ligand binding activity values of GPCR 'G' w.r.t ~1000.

$$\hat{a}(G) = \bar{a}(G) + \bar{l}(p_2(G) - p_1(G))$$

2.4.3 Discussion

Like 'Correlated distances' method, 'Correlation properties' doesn't use any clustering and therefore it is more robust than the 'Hierarchical clustering' method.

In addition, the results of 'Correlated properties' method are more detailed. They include ligand property specification in addition to the position and amino-acid property. This addition can be very helpful, especially for the main industrial consumer of this data – drug companies. Drug companies search for drugs that bind strongly (have high affinity) with specific GPCRs. Given a list of triplets (output of the algorithm) and a GPCR of interest, it is possible to retrieve the amino-acid in the position specified by the triplet and it's amino-acid property value. If the value predicts a strong correlation with the ligand property specified by the triplet, the search for drugs can be narrowed to ligands with specific values w.r.t the ligand property. If not, results from the next iteration should be used.

The main weakness of this method is related to the neutralization procedure. The procedure fits the data to two regression lines, calculated one upon the other (the y-values fitted by the second regression line are the slopes found for the set of the first regression lines). Every such fit incorporates within it error and noise, and therefore the adjustment of the ligand binding activity data-set is expected to decrease the accuracy of the data. From one iteration to the next, the amount of decrease in accuracy is increased.

2.5 Incorporation of Two-way clustering

2.5.1 Introduction

'Two-way clustering' uses Gene expression clustering algorithms to find triplets of position, amino-acid property and ligand property that affect binding.

Any clustering algorithm that finds sub-groups of both genes and conditions can be used. In this work we used 'Iterative signature algorithm' [10] .

Applied upon the ligand binding activity data set, a cluster would include a subgroup of GPCRs that show similar ligand binding activity values w.r.t a subgroup of ligands. Returning to the aminergic GPCRs' binding-site example – we would expect to find a cluster of all GPCRs that have Aspartic acid (D) in position 689 and all charged ligands. Once such a cluster is found, identifying position 689 and amino-acid property 'Charge' can be done by any of the 3 methods above;

'Hierarchical clustering' method would find position and amino-acid property combinations such that the null hypothesis, that the distributions of the values for GPCRs within the cluster vs. GPCRs outside the cluster are equal, would yield the smallest p-value. A similar method can be applied to find ligands' properties, such that the ligand property values w.r.t ligands within the cluster vs. ligands outside the cluster are the most distant.

'Correlated distances' method would be applied to each cluster separately instead of the entire set. The probability of finding the correct position and property is expected to improve due to the disposal of GPCRs and ligands that add noise to the data. In addition, combining 'Correlated distances' with 'Two-way clustering' enables finding 'second order' results (position-property combinations that affect only a small group of

the GPCRs, such as GPCRs with uncharged amino-acid in position 689) without relying on the initial results of ligand specificity-determining positions and amino-acid properties as mediators .

Similarly, 'Correlated properties' would be applied to each cluster separately.

2.5.2 The method

Activate 'Iterative signature algorithm' [10] on the ligand binding activity data set (or ligand-properties correlations data set). The algorithm's output includes 'modules', containing subgroups of GPCRs and ligands (or ligand properties).

These modules can be incorporated into the three methods in different ways to improve their expected results, as explained below.

'Hierarchical clustering' method

1. Activate a modified version of 'Hierarchical clustering' method. Instead of splits in a dendrogram, the clusters of GPCRs created by 'Iterative signature algorithm' will be analyzed vs. their complementary subgroups.
2. It is possible to activate a second modified version of "hierarchical clustering', to find ligand properties that affect binding. Instead of splits in the dendrogram, the clusters of ligands will be analyzed vs. their complementary subgroups. Instead of amino-acid properties retrieved for each position separately, ligand properties would be used.

'Correlated distances' method

1. Activate 'Correlated distances' method once for each module, using only the GPCRs and ligands included in the module, to find combinations of position and amino-acid property that affect binding specificity in each module.

   In order to identify positions and properties that differentiate between the modules, the original ligand binding activity data set should be defined as a module as well.
2. It is possible to modify 'Correlated distances' method to find ligand properties that affect binding in each module. Again, the procedure will be applied once to each module using only the GPCRs and ligands included in the module. The distances between ligand binding activity values would be calculated between the columns of the ligand binding activity matrix (each column represents ligand binding activity values of all GPCRs and one ligand) and not rows as in the original procedure. The properties distances would be calculated upon the ligands' properties dataset (instead of amino-acid properties, specific to each position, as in the original procedure).

'Correlated properties' method

Activate the original 'Correlated properties' method once for each module, using only the GPCRs and ligands included in the module.

Like in 'Correlated distances' method, in order to identify positions and properties that differentiate between the modules, the original ligand binding activity data set should be defined as a module as well.

2.5.3 Discussion

Combining Two-way clustering with the various methods brings significant improvements to each one of them;

The 'Hierarchical clustering' method's main disadvantage, its sensitivity to the hierarchical clustering structure, is solved due to the flexible structure of the clustering algorithm. The rigidity of the hierarchical clustering is due to the forced tree structure - wrong classification in a split results in wrong clustering of the entire sub-tree, while clustering algorithms find various clusters, possibly overlapping.

'Correlated distances' method lacks the ability to find position-property combinations affecting only sub-group of the GPCRs (unless it is induces by the best scoring position-property combination found in previous iteration). It is clear that a descent in the percentage of GPCRs affected by the position-property combination would yield a descent in the score associated with it.

Combining 'Two-way clustering' cancels this disadvantage by allowing the activation of 'Correlated distances' method on each GPCRs group separately. This improvement allows finding of additional 'second order' combinations that take affect when the major combination is not active (for example if the specified property is 'Charge' and the ligand is not charged).

Similar improvements apply to combining 'Two-way clustering' with 'Correlated properties' method. The 'Correlated properties' method solves the inability of finding second order results by repetition of the analysis after neutralizing the affect of the found major results. Nevertheless, as discussed in chapter 2.4.3, the neutralization procedure is highly exposed to error and noise. In fact, Two-way clustering offers highly-effective improved alternative to the neutralization procedure used in 'Correlated properties' method.

2.6 Results analysis

After a list of candidate positions and properties is created by one or few of the suggested methods, two analyses were performed: the significant transmembranal positions were superimposed on the resolved 3D structure of rhodopsin and analyzed by kPROT [11]. kPROT (Knowledge-based Scale for the Propensity of Residue Orientation in Transmembrane Segments) predicts the angular orientation of the Transmembrane helices by calculating for each residue the ratio of its proportions in single and multiple TM spans among a set of 5000 non-redundant protein sequences. It is based on the assumption that residues that tend to be exposed to the membrane are more frequent in TM segments of single-span proteins, while residues that prefer to be buried in the transmembrane bundle interior are present mainly in multi-span TMs.

Superimposing the significant positions on the 3D structure of rhodopsin allows estimating the proximity of the positions to the extracellular side of the receptors, the proximity between the positions and their general positioning. The alignment of the transmembranal regions of the GPCRs to rhodopsin expected to be highly accurate due to the high conservation of the transmembranal segments.

Chapter 3
Results

3.1 Introduction

Each of the methods detailed above yields a slightly different form of results ,but they all specify pairs of position in the alignment and amino-acid property that are predicted to affect binding.

One indication for the correctness of the results is the finding of the highly conserved Asp position in the third transmembranal region, which is known to be the binding site of aminergic GPCRs, coupled to the amino-acid property 'Charge'. Additional indications are the angular orientation of the found binding site, as predicted by kPROT [11], and the 3D position of the found residues as superimposed on the resolved 3D structure of rhodopsin. We assume that binding residues are localized in the plane that faces the interior of the protein (as opposed to the membrane) and closer to the extracellular end of the GPCR than to the intracellular end.

Due to the similar structure of all GPCRs and to the high conservation of the transmembranal regions in GPCRs, it is possible to identify the matching position in the GPCRs included in the input data sets with relatively high accuracy. This position is numbered 689.

All the following results were retrieved by applying the methods only to the transmembranal regions of the GPCRs.

3.2 Method 1 – Hierarchical clustering

The output of method 1 includes lists of position and amino-acid property combinations. A separate list is created for each split in the dendrogram (that was determined to be stable by the bootstrap analysis).

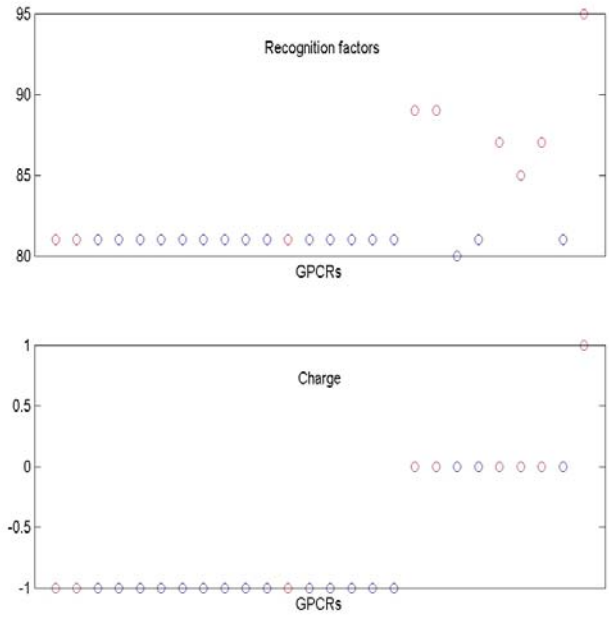
Figure 4 presents the results of applying 'Hierarchical clustering method' to the ligand properties binding profiles (details in chapter 2.1.4- Ligands properties) and the clusters of amino-acid properties (details in chapter 2.1.2 - Amino-acid properties).

For all splits, the FDR-adjusted significance threshold is 0, and no significant results were found. Nevertheless, the position-property combinations that got the highest scores were examined. Figure 5 presents the best-scoring combination for split 1 – the root of the dendrogram (the score is equal to the p-value calculated by the two-sample Kolmogorov-Smirnov test). It matches exactly the known aminergic binding site and amino-acid property 'charge' (one of the two amino-acid properties included in amino-acid properties cluster 17, which was selected by the procedure).

Figure 5(A) shows the amino-acids found in the position selected by the procedure, divided according to the classification of GPCRs by the dendrogram. It can be seen

that the classification of GPCRs by the first split in the hierarchical clustering is not optimal – D is the only negatively charged amino-acid, and therefore we would expect to see GPCRs with D separated from all other GPCRs. Figure 5(B) shows the values of the two amino-acid properties included in the selected cluster – 'Recognition Factors' and 'Charge', w.r.t the amino-acids. GPCRs in different sides of the split are colored with different colors. It can be seen that for the GPCRs in the top split (most of which have amino-acid 'D' in the selected position), the values of 'Recognition factors' are equal to ~81 while for the GPCRs in the bottom split the values vary from 81 to 95. The 'Charge' property is also generally lower for GPCRs in the top split than ones in the bottom split. Figure 5(C) presents KPROT analysis for the selected position. It is reasonable to assume that binding-sites face the gorge – the interior hole in the GPCR to which the ligand enters, and not the membrane. kPROT analysis predicts the angular orientation of the helix, and therefore, according to the location of the position within the helix, predicts whether it faces the membrane or the gorge. The red thick line is directed to the predicted direction of the membrane, and therefore binding sites are supposed to appear in the opposite direction. It can be seen that the selected position, marked by a red 'D', is indeed facing the opposite direction. Figure 5(D) shows the position in rhodopsin that matches the best-scoring position (by alignment of the transmembrane region) - colored in magenta. The top of the image is the extracellular milieu, and the bottom – the intracellular. Binding sites are predicted to be located closer to the extracellular side. It can be seen that the best-scoring position applies to this condition as well.

A **B**



C **D**

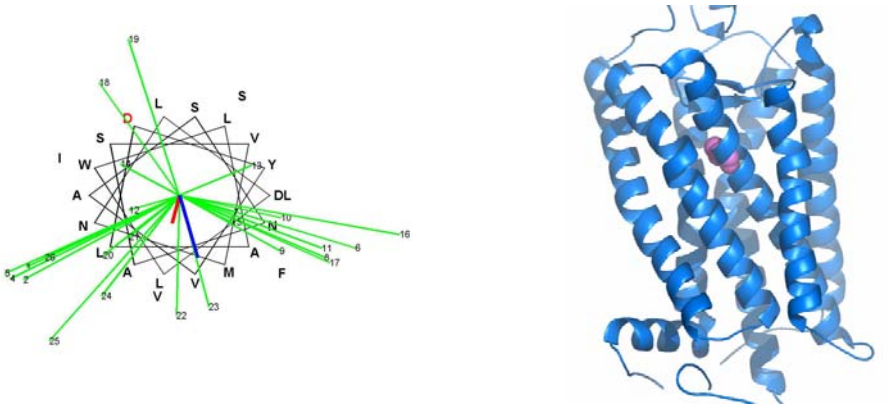
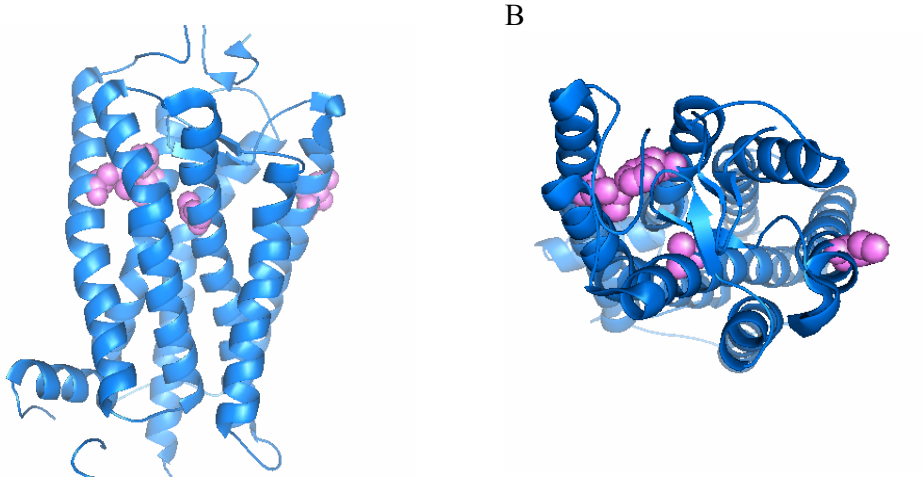


Figure 5: Best scoring position and amino-acid-properties-cluster for the main split in the hierarchical clustering tree. This result (both the position and the amino-acid property 'Charge') coincide exactly with the aminergic binding site. **(A)** The alignment of the best-scoring position. The two groups of GPCRs defined by the main split in the Hierarchical clustering (shown in figure 4.C) are surrounded by red rectangles**. (B)** The values of the amino-acid properties included in the best-scoring amino-acid-properties-cluster: Recognition factors and Charge. GPCRs in different sides of the split are colored with different colors. **(C)** kPROT analysis for transmembrane region #3. The best-scoring position is shown as a red 'D'. The red thick line is the direction predicted to face the membrane. It is equal to the mean of the directions predicted to face the membrane calculated for each GPCR in the alignment separately (plotted as green lines). The blue line directs to the position, in which the amino-acids show the highest variability. The wide angle (~120°) between the red line and the selected position indicates that it faces the interior of the protein, which is the area predicted to bind ligands. **(D)** Structure of bovine rhodopsin. The position in rhodopsin that matches to the best-scoring position (by alignment of the transmembrane region) is colored in magenta.

The location of the 4 best scoring positions for split 1 is presented in Figure 6. The figure shows that all positions cluster around the aminergic binding site. Two of the four positions face the interior of the protein (including the best scoring position, presented in figure 5), and the other two face ~90° and ~30° from the membrane, and connect between helices 5 and 7.

A B



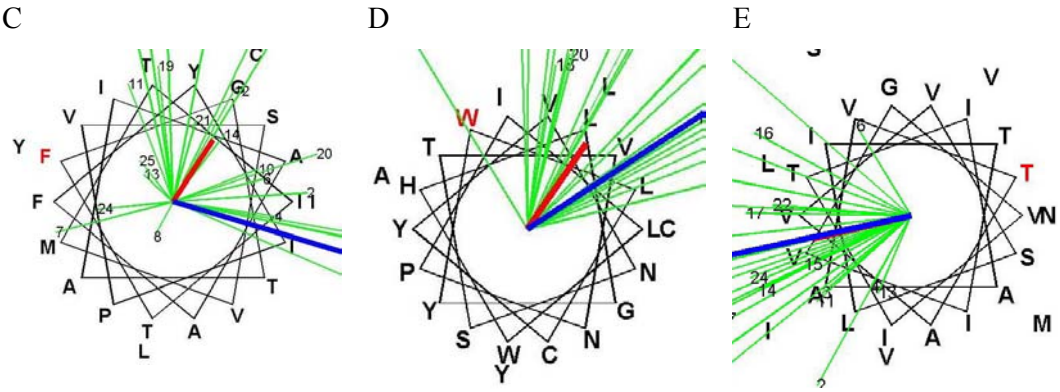


Figure 6: Four best-scoring positions for the main split in the hierarchical clustering tree, marked on the structure of bovine Rhodopsin (A, B) and their kPROT analysis (C,D,E) (not including the best scoring position, shown in Figure 5(C)). **(A)** A side view as seen from within the membrane. **(B)** A view from the extracellular milieu. **(C)** kPROT analysis for the 5th transmembrane helix. The 3rd position is marked as a red 'F', and predicted to face the 90° from the membrane. **(D)** kPROT analysis for the 7th transmembrane helix. The 7th position is marked as a red 'W' and predicted to face 30° from the membrane. **(E)** kPROT analysis for the 1st transmembrane helix. The 8th position is marked as a red 'T' and predicted to face the gorge.

The best scoring results for the first 3 splits and their scores are detailed in Table 1. The dendrogram that yielded the results is shown in Figure 7. The first three splits, for which results are shown in the table below, are colored red.

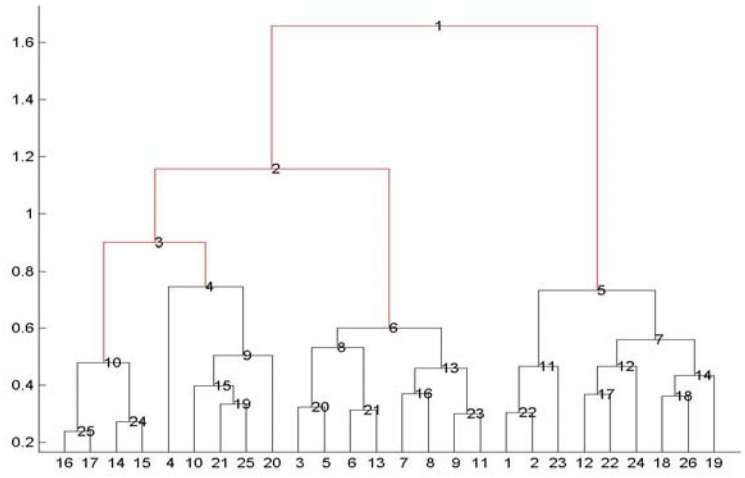


Figure 7: Hierarchical clustering of the 26 GPCRs. The distances between the GPCRs used for the creation of the dendrogram are Euclidean distances. Each GPCR is represented by a vector of size 164, where the i'th elements is the Peasron correlation between the ligand binding activity values of the GPCR and the values of the i'th ligand property w.r.t to the 1068 ligands. The dendrogram is shown also in Figure 4(B). The first three splits, for which results are shown in the table below, are colored red and the splits' numbers are presented.

Split	Position	Amino-acid properties	P-value
1	689 (3,7)	Cluster 17 (Recognition factors; Charge)	0.0051
2	1255 (6,10)	Clusters 1 11 14 16 (18 amino-acid properties)	0.0048
	650(2,15)	Clusters 2 3 5 6 9 12 (21 amino-acid properties)	0.0048
	683(3,1)	Cluster 18 (Overall amino acid composition; Number of codon(s); Amino acid composition (%) in the Swiss-Prot Protein Sequence data bank)	0.0048
3	605 (1,24)	Clusters 2 15 (Retention coefficient in HFBA; Retention coefficient in TFA; Antigenicity value X 10; polarity)	0.0069
	583 (1,3)	Cluster 3 (Optimized matching hydrophobicity; Mobilities of amino acids on chromatography paper (RF))	0.0069
	1304 (7,6)	Clusters 6 15 (Molecular weight; Refractivity; Average area buried on transfer from standard state to folded protein; Antigenicity value X 10; polarity)	0.0069
	1266 (6,20)	Clusters 11 15 16 18	0.0069
	655 (2,20)	Cluster 14	0.0069
	1004 (5,19)	Cluster 14	0.0069
	747 (4,7)	Cluster 15 (Antigenicity value X 10; polarity)	0.0069
	683 (3,1)	Cluster 18 (Overall amino acid composition; Number of codon(s); Amino acid composition (%) in the Swiss-Prot Protein Sequence data bank)	0.0069

Table 1: Best scoring position & amino-acid properties combinations. The position indices represent the location of the position in the original input sequence alignment. The first number in the parentheses is the index of the transmembranal region, to which the position belongs. The second number is the index of the position within the transmembranal region. The second column specifies the indices of the amino-acid property clusters that were found along with the positions. For most of the clusters the amino-acid properties' names are specified as well. The third column is the p-value associated with two-sample Kolmogorov-Smirnov test, applied on the values of the specified amino-acid properties cluster and position, where the first distribution relates to GPCRs in the left side of the root split in the dendrograms and the second – to the right side.

3.3 Method 2 – Correlated distances

The output of 'Correlated distances' method includes pairs of position and amino-acid property, such that the pairwise distances between the GPCRs, calculated upon the value of the amino-acid property w.r.t the amino-acids in the position, are highly correlated with the pairwise distances between the GPCRs, calculated upon their ligand binding activity values. The scores associated to the pairs are p-values, calculated by 10,000 random shuffles of the data.

'Correlated distances' was applied to the normalized ligand binding activity values and amino-acid properties clusters. The first iteration – created by the original group of 26 GPCRs, yielded only one significant result (p-value <= 0.05 FDR corrected), that matches exactly the binding site of aminergic GPCRs and the known amino-acid property – 'Charge' (clustered with an additional amino-acid property – 'Recognition

factors'). The correlation coefficient for that position and amino-acid property is 0.4231 (vectors of size 325) and the p-value is 0, meaning that 10,000 shuffles of the data didn't yield an equal or better result.

Figure 8 plots the two correlated vectors that yielded the best score. The x-axis is pairwise distances between the values of amino-acid properties cluster 17 w.r.t the amino-acids of the 26 GPCRs in position 689. The y-axis is pairwise Euclidean distances between the normalized ligand binding activity values of the 26 GPCRs. It can be seen there is a strong positive correlation between them.

kPROT analysis and the location of the selected position superimposed on bovine-rhodopsin are shown in Figure 5(A-B).

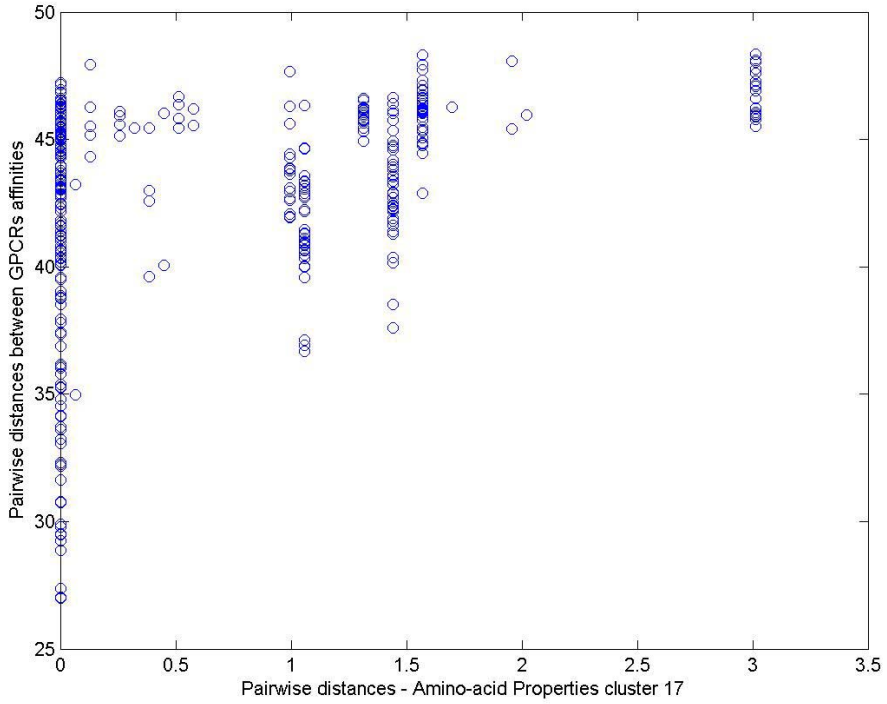


Figure 8: The best-scoring, significant result of the activation of 'Correlated distances' method upon the GPCRs' normalized ligand binding activity data set and the amino-acid properties clusters data set. Only one pair was found significant – position 689 (in transmembranal helix 3) and amino-acid properties cluster-17, which groups two amino-acid properties – 'Charge' and 'Recognition factors'. The values of cluster-17 are equal to the mean of the normalized values of 'Charge' and 'Recognition factors' w.r.t the 20 amino-acids. Position 689 is known to be the aminergic binding site, and amino-acid property 'Charge' is known to affect amine binding. The X-axis of the plot is pairwise distances between the values of cluster-17 w.r.t the amino-acid of the 26 GPCRs in position 689. The y-axis is pairwise Euclidean distances between the normalized ligand binding activities of the 26 GPCRs. The Pearson correlation coefficient between these two vectors is 0.4231, and the associated p-value, measured by 10,000 random-shuffles of the data, is 0.

The first iteration produced a division of the 26 GPCRs into two groups such that bio-amine receptors were perfectly separated from non bio-amine. The two groups were the inputs for the second iteration.

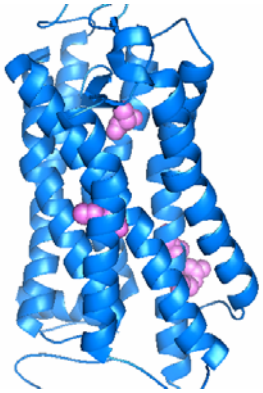
In the second iteration eight significant results (that include four different residues) were found for the non bio-amine group and none for the bio-amine group. The results for the non bio-amine receptors are shown in Table 2.

kPROT analysis for the four positions and their location superimposed on bovine-rhodopsin are shown in Figure 9.

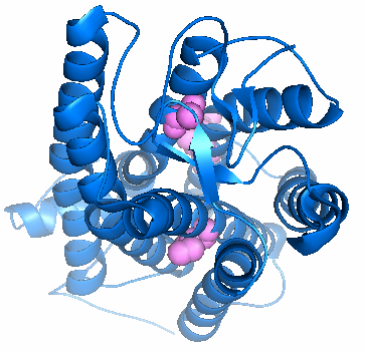
Position	Amino-acid properties	P-value
695 (3,13)	Cluster 17 (recognition factors; charge)	0.0001
1248 (6,3)	Cluster 3 (Optimized matching hydrophobicity; Mobilities of amino acids on chromatography paper (RF))	0
	Cluster 5 (polarity and four hydrophobicity scales)	0.0001
	Cluster 8 (Retention coefficient in HPLC, pH 2.1/pH 7.4)	0.0001
	Cluster 9 (seven hydrophobicity related properties)	0.0001
1252(6,7)	Cluster 8 (Retention coefficient in HPLC, pH 2.1/pH 7.4)	0.0001
	Cluster 15 (Antigenicity value X 10; polarity)	0.0001
1304(7,6)	Cluster 16 (Atomic weight ratio of hetero elements in end group to C in side chain; Conformational parameter for beta-turn; Normalized frequency for beta-turn; Conformational parameter for beta-turn (computed from 29 proteins); Conformational parameter for coil)	0

Table 2: Significant position & amino-acid properties cluster combinations found by the second iteration of 'Correlated distances' method, applied to the non bio-amine receptors group. The position indices represent the location of the position in the original input sequence alignment. The first number in the parentheses is the index of the transmembranal region, to which the position belongs. The second number is the index of the position within the transmembranal region. The second column specifies the indices of the amino-acid properties clusters that were found along with the positions, followed by the amino-acid properties included in them). The third column is the p-value, estimated as the percent of 10,000 random shuffles that yield a higher or equal correlation.

A



B



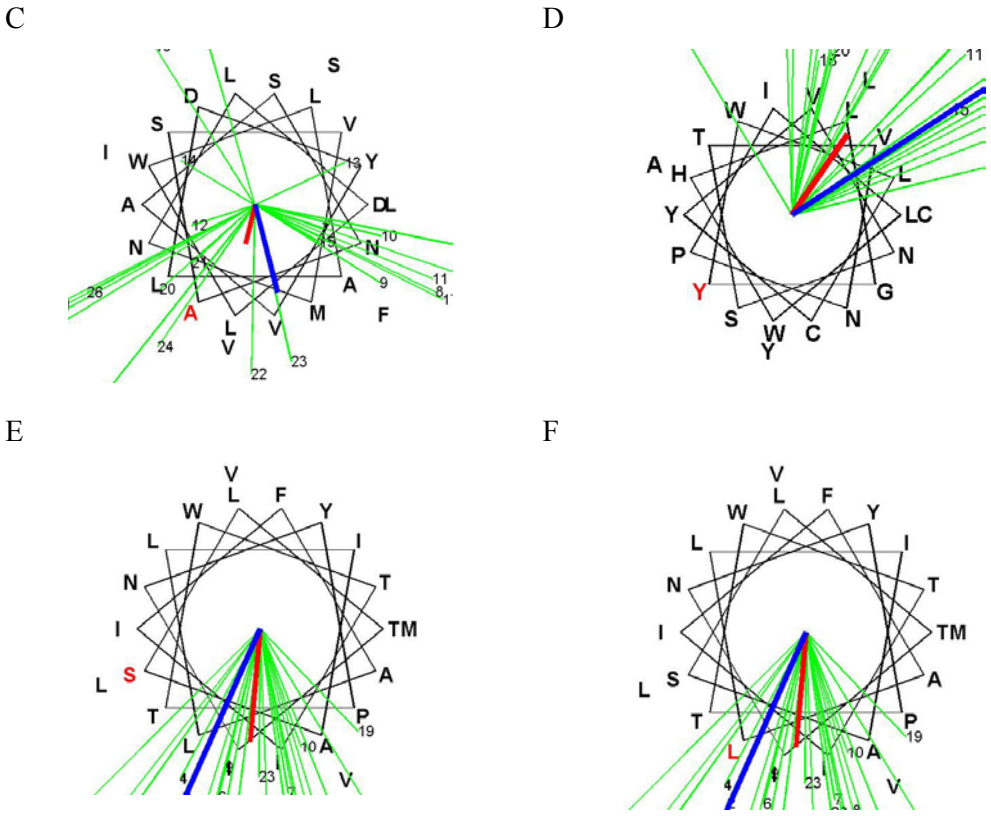
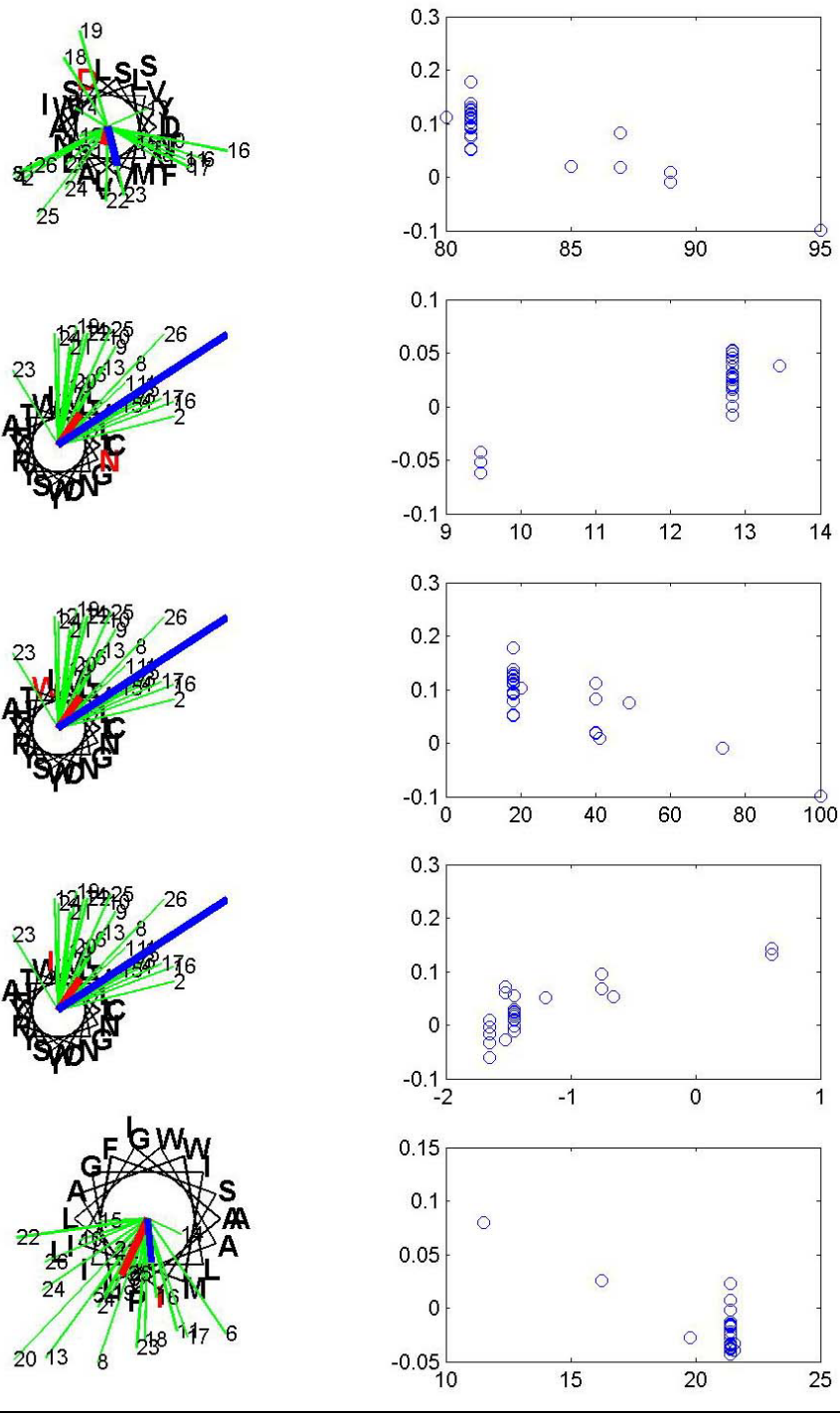


Figure 9: Structural analysis of the four significant residues found by the second iteration of 'Correlated distances' method – relate only to the nine non-bio-amine receptors. (A, B) show the location of the residues on the structure of bovine Rhodopsin. (C-F) is the residues' kPROT analysis. (A) A side view as seen from within the membrane. (B) A view from the extracellular milieu. (C) kPROT analysis for position 695 from the 3rd transmembrane helix. The position is marked as a red 'A', and predicted to face the membrane. (D) kPROT analysis for position 1304 from the 7th transmembrane helix. The position is marked as a red 'Y' and predicted to face the gorge. (E) kPROT analysis for position 1248 from the 6th transmembrane helix. The position is marked as a red 'S' and predicted to face 60° from the membrane. (F) kPROT analysis for position 1252 from the 6th transmembranal helix. The position is marked as a red 'L' and predicted to face the membrane.

3.4 Method 3 – Correlated properties

The output of 'Correlation properties' method includes triplets of position, amino-acid property and ligand property that show high correlation. The triplets presented below are the results of applying the method upon the original amino-acid properties (not the clustered properties). 'Correlated properties' method, like 'Hierarchical clustering' method, assigned the best score to the position that matches the aminergic binding site, but with the amino-acid property 'Recognition factors' and not the property known to affect binding – 'Charge'. It is not known whether the found ligand-property, Isis_127, is related to charge. A total of five positions (in 12 triplets) were found significant, with p-values <= 0.05 FDR corrected. Two of these positions match the best scoring position found by the 'Hierarchical clustering' method for the root split. The second iteration (applied to adjusted ligand binding activity values, after the effect of the triplets found in the first iteration were neutralized), no significant results were found.

A



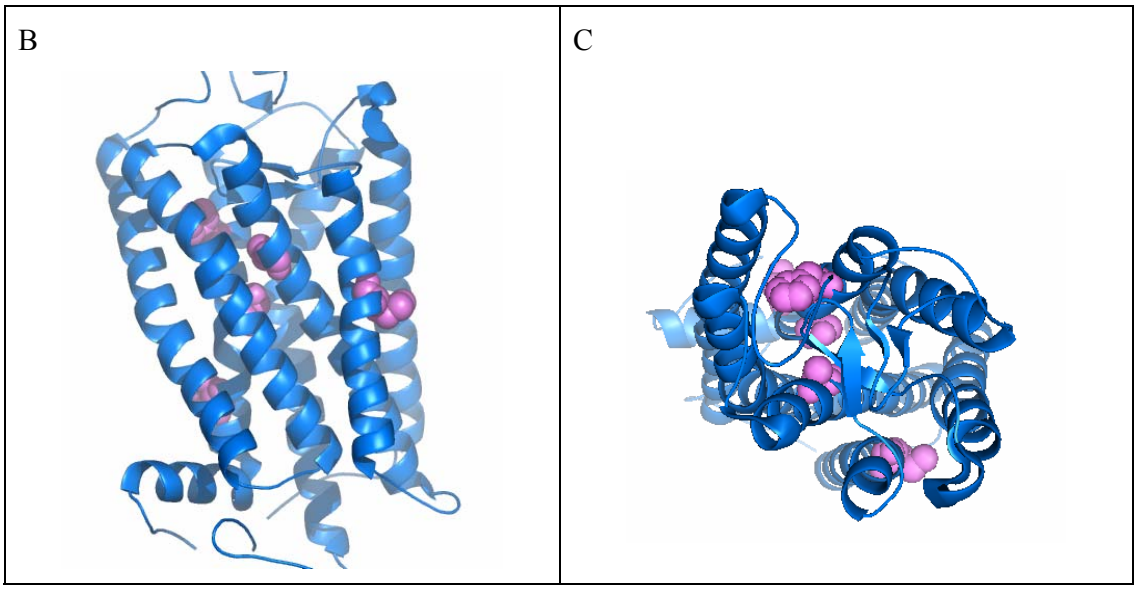


Figure 10: Results of the 'Correlated properties' method. The algorithm found 12 combinations of position, amino-acid-property and ligand-property with significant p-values. The combinations include 5 different positions. The best scoring position coincides exactly with the aminergic binding site, but not with the known amino-acid property. **(A)** For each of the 5 selected positions, two figures are shown: (A.Left) The KPROT analysis for the selected position. The position is colored in red. The red thick line is the direction predicted to face the membrane. (A.Right) A plot of the correlation between the GPCRs' ligand binding activity and the values of the selected ligand-property w.r.t 1068 ligands (Y-axis) against the values of the selected amino-acid property in the selected position (X-axis). For example, the top plot presents the combination: position 689 (7^{th} in transmembrane region 3), amino-acid property 'Recognition factors' and ligand-property 'isis_125'. It shows that a low value for the amino-acid property 'Recognition Factors' leads to a strong dependency between the measured activity of a GPCRs following the binding of a ligand and the value of the ligand's property 'isis_125'. Although position 1310 (presented in the second plot) was found to be significant with 8 different amino-acid properties, only one property is plotted. The full list of results is shown in table 2. **(B)** The 5 selected positions marked on the structure of bovine rhodopsin, as seen from within the membrane. **(C)** The 5 selected positions as seen from the extracellular milieu.

Position	Amino-acid property	Ligand Property	P-value
689 (3,7)	Recognition factors	Isis_125	2.05e-008
1310 (7,12)	Bulkiness	Isis_163	2.14e-008
	Normalized frequency for beta-sheet		2.16e-008
	Mobilities of AAas on chromatography paper (RF)		2.29e-008
	Number of codon(s)		2.34e-008
	Recognition factors.		3.75e-008
	Molar fraction (%) of 2001 buried residues		1.13e-007
	Retention coefficient in HPLC, pH 2.1		1.72e-007
	Conformational parameter for alpha helix (computed from 29 proteins)		1.95e-007
1305 (7,7)	Relative mutability of amino acids (Ala=100)	Isis_125	2.36e-007

Position	Amino-acid property	Ligand Property	P-value
1316 (7,18)	Hydrophobicity (free energy of transfer to surface)	Isis_93	3.81e-007
751 (4,11)	Bulkiness	Isis_60/73	4.13e-007

Table 3: Results of the 'Correlated properties' method (first iteration) – triplets of position, amino-acid property and ligand property that were found significant (p-value <= 0.05 FDR corrected).

3.5 Method 4 - Incorporation of Two-way clustering

Iterative signature algorithm was applied to the ligand binding activity data set, after normalization and outliers elimination. Iterative signature algorithm was designed to cluster gene chips data, and therefore the input ligand binding activity data set was formatted such that GPCRs will be treated as genes and ligands - as conditions.

The algorithm yielded two results. With threshold=0.1 one module was identified, that includes 14 GPCRs and 687 ligands. The module's GPCRs include 13 bio-amine binding GPCRs (out of 17 in the original set) and one ATP/ADP binding GPCR (out of one in the original data set). None of the neuropeptide binding GPCRs were included in the cluster (6 in the original data set).

With threshold=0.3 three modules were identified. The first contains 9 bio-amine binding GPCRs and 617 ligands; the second contains 6 GPCRs and 694 ligands. Five of the GPCRs are bio-amine (including one that was selected in the first module as well) and the 6^{th} is the ATP/ADP binding GPCR. The third contains 3 GPCRs and 664 ligands. All three selected GPCRs bind neuropeptides (gastrin and cholecystokinin / galanin / neuropeptide Y).

Neither 'Correlated distances' method nor 'Correlated properties' method, applied to these modules, found any significant results (defined as results with p-values lower or equal to the significance threshold calculated by FDR with 0.05). However, some of the methods, when applied to the module of 14 GPCRs and 689 ligands (threshold 0.1), found significant results according to differently calculated significance thresholds.

'Correlated distances' method, applied to this module, identified two specificity-determining residues and four amino-acid properties clusters with p-values < 0.1 (Bonferroni corrected), detailed in Table 4.

The residues are not predicted to face the interior of the protein by kPROT.

Position	Amino-acid properties	P-value
683 (3,1)	Cluster 3 (Optimized matching hydrophobicity; Mobilities of amino acids on chromatography paper (RF))	2.0e-4
	Cluster 7 (Normalized consensus hydrophobicity scale; Hydration potential)	5.0e-4
	Cluster 18 (Overall amino acid composition; Number of codon(s); Amino acid composition (%) in the Swiss-Prot Protein Sequence data bank)	4.0e-4

Position	Amino-acid properties	P-value
988 (5,4)	Cluster 19 (Conformational parameter for alpha helix; Normalized frequency for alpha helix; Conformational parameter for alpha helix computed from 29 proteins).	5.0e-4

Table 4: Results of the 'Correlated distances' method applied to a module found by 'Iterative signature algorithm'. The module consists of 13 bio-amine GPCRs and one ATP/ADP GPCR.

The 'Correlated properties' method found four results with p-values < 1 (Bonferroni corrected), detailed in Table 5. Three of the four are predicted by kPROT to face the interior of the protein.

Position	Amino-acid properties cluster	Ligand Property	P-value
761 (4,21)	Cluster 19 (Conformational parameter for alpha helix; Normalized frequency for alpha helix; Conformational parameter for alpha helix computed from 29 proteins).	Isis_3	3.2844e-7
745 (4,5)	Cluster 19 (See previous row)	Isis_68	1.6761e-6
1247 (6,2)	Cluster 7 (Normalized consensus hydrophobicity scale; Hydration potential).	Isis_130	9.5227e-7
603 (1,23)	Cluster 17 (Recognition factors; Charge).	Isis_130	6.0524e-7

Table 5: Results of the 'Correlated properties' method applied to a module found by 'Iterative signature algorithm'. The module consists of 13 bio-amine GPCRs and one ATP/ADP GPCR.

3.6 The results – Discussion

All the methods discussed above were expected to identify combinations of residues and amino-acid properties that affect binding (in 'Correlated properties' method – combinations of residues, amino-acid properties and ligand properties).

The success of all the methods depends on a basic assumption, that GPCRs of the same type (where the type can be general, as aminergic receptors, or specific as dopamine receptors) have similar ligand binding activity.

This assumption is only partially realized in the data-set, as demonstrated by Figure 11 and Figure 12 (generated by CTWC [13]). Figure 11 shows the normalized ligand binding activity values of the 26 GPCRs, reordered by similarity (both GPCRs and ligands). Figure 12 shows the GPCRs' distances matrix.

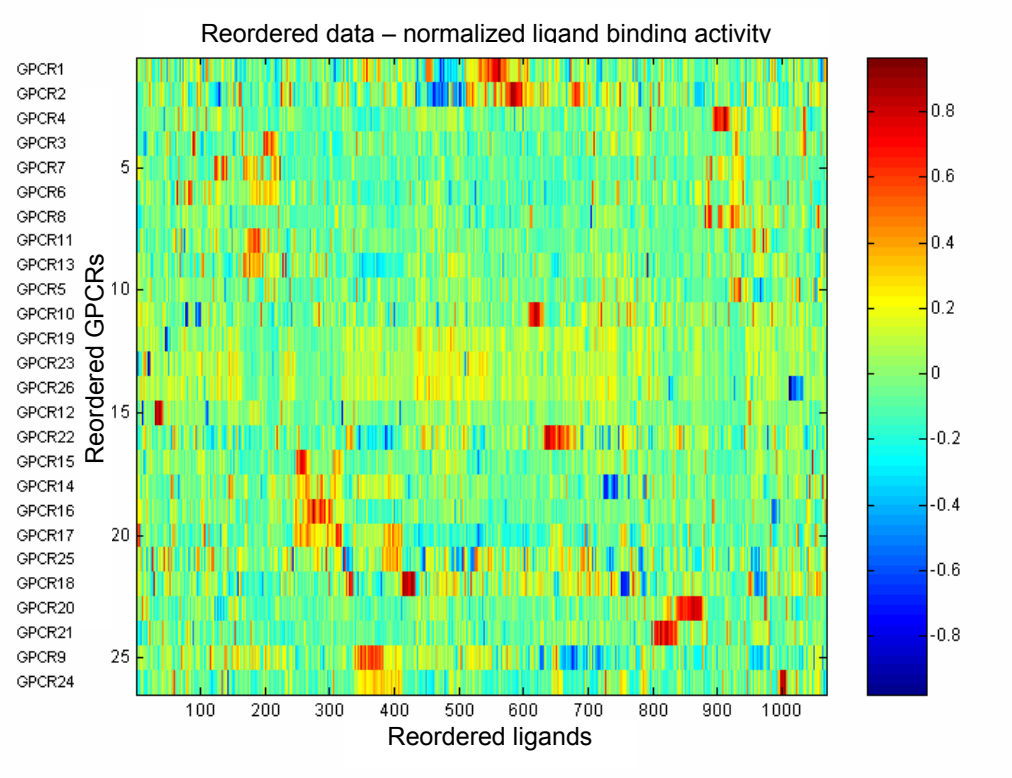


Figure 11: The normalized ligand binding activity of the 26 GPCRs, reordered (both GPCRs and ligands) by their similarities. The reordering and creation of the image were done by CTWC.

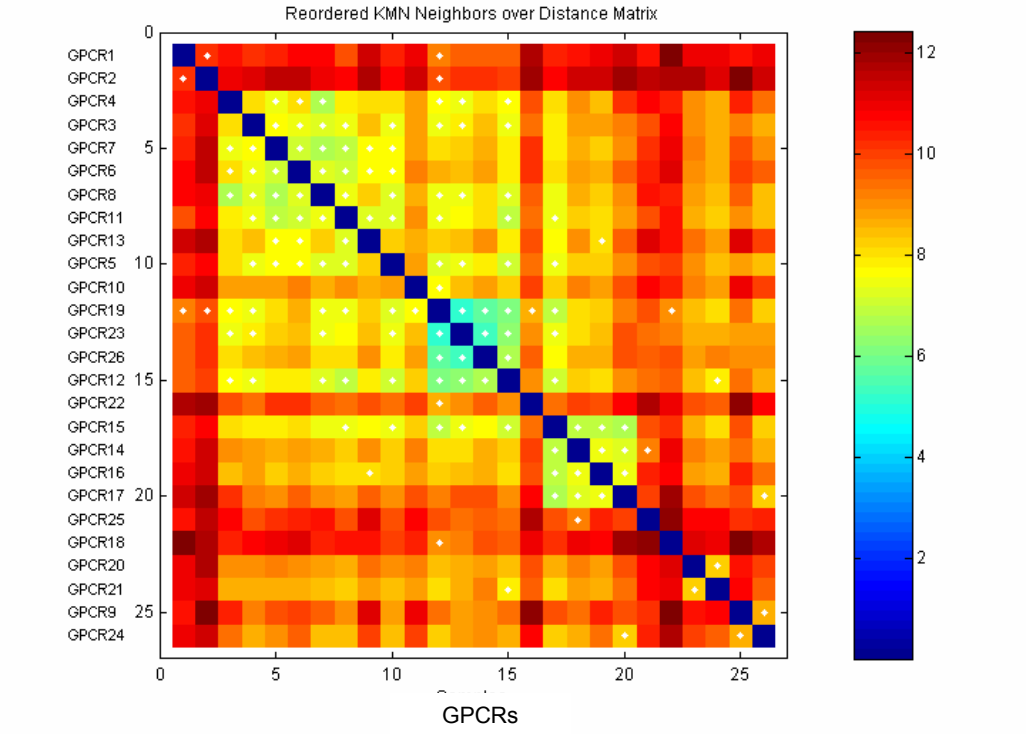


Figure 12: Pairwise distances between the normalized ligand binding activity of the 26 GPCRs, reordered by their similarities. The reordering and creation of the image were done by the CTWC.

For example, GPCRs 1-5 are Muscarinic acetylcholine receptors. GPCRs 9 and 15 are both histamine receptors. Nevertheless, both groups do not have similar binding profiles. On the other hand, GPCRs 19, 23 and 26, that show similarity, bind different ligands (19 – adenosine, 23 – angiotensin and 26 – neurotensin) two of which are peptides and one ribonucleotide (adenosine).

These deviations from the basic assumption can happen due to various reasons. However, it is likely that the identity of ligands in the data-set has great influence. For example, if histamine is not one of the ligands, then it is probable that histamine receptors do not show exceptional binding similarity.

The only result that was expected in advance is position 689 – experiments confirm that bio-amine ligands are attached to the GPCR by a charged residue in this position. Despite the deviation from the first assumption, as mentioned above, residue 689 was identified by all methods. 'Correlated distances' identified it perfectly as the only significant result, 'Correlated properties' identified it along with four other residues and 'Hierarchical clustering' did not mark it as significant but assigned it the lowest p-value (out of more than 3000 options). While 'Correlated distances' method identified the aminergic binding site directly from the ligand binding activity data set, the other two methods rely on an additional data set of 164 ligand properties. The alternative binding data set includes correlations between the ligand binding activity of each GPCR and the values of each ligand property, as shown in Figure 13 and Figure 14. Figure 13 presents the correlations of the 26 GPCRs, reordered by similarity (both GPCRs and ligand properties). Figure 14 shows the GPCRs' distances matrix.

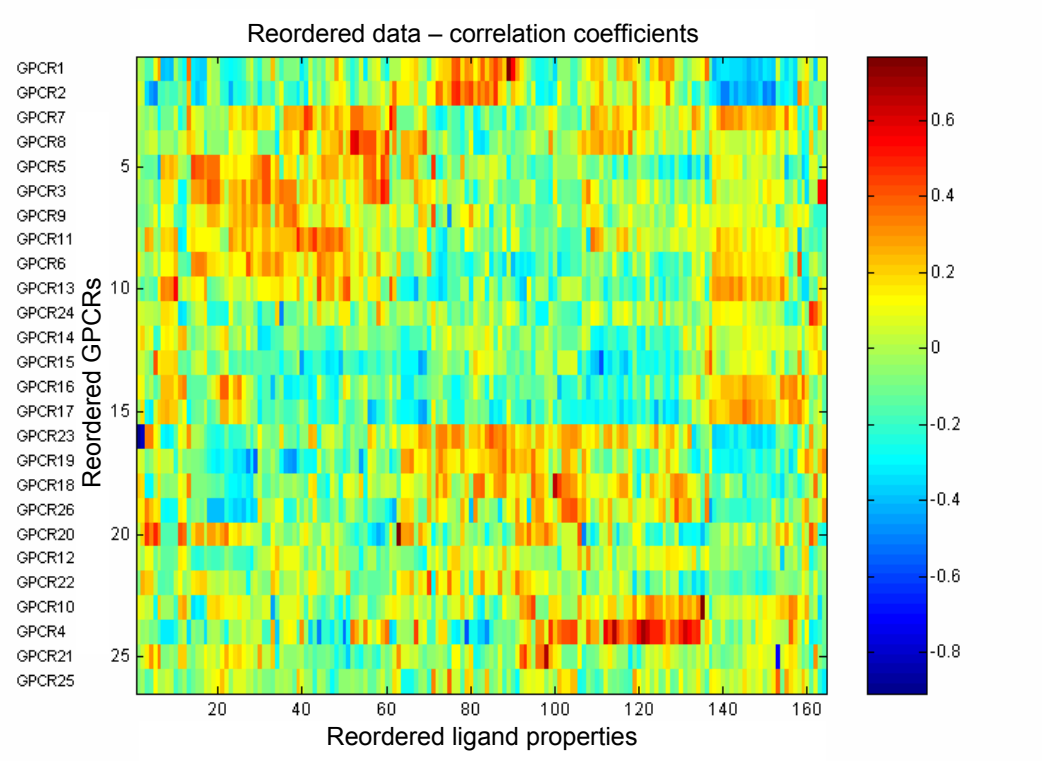


Figure 13: Correlation coefficients of the 26 GPCRs w.r.t 164 ligand properties, reordered (both GPCRs and ligand properties) by their similarities.

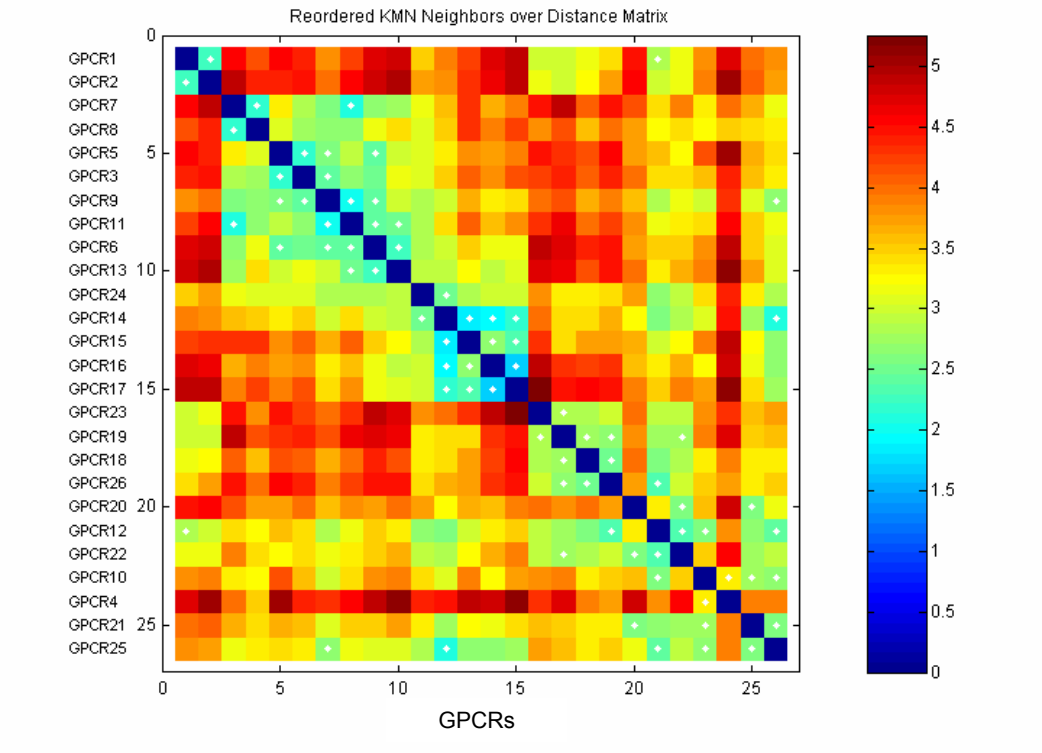


Figure 14: The pairwise distances between the correlations of the 26 GPCRs, reordered by their similarities. The reordering and creation of the image were done by the CTWC algorithm.

Coupled Two-Way Clustering, when applied to the ligand-properties correlation coefficients data-set, was more successful in discriminating between bio-amine receptors and the rest. GPCRs 1-17 are aminergic receptors, GPCRs 18-19 bind adenosine (ribonucleotide), GPCR 25 binds ATP/ADP and the rest (GPCRs 20-24 and GPCR 26) bind various peptides. It can be seen that all non-aminergic receptors (GPCRs 18-26) except for GPCR 24 are adjacent to each other (in the bottom of Figure 13 and Figure 14) along with three aminergic receptors – GPCR 12, GPCR 10 and GPCR 4.

The hierarchical clustering method applied to the ligand-properties-correlations data set also divided the GPCRs into two groups such that aminergic receptors were separated from the rest of the GPCRs with slight deviations, as shown in Figure 7. The deviations include three aminergic receptors (out of 17) classified into the non-aminergic group (GPCRs 1, 2 and 12) and three non-aminergic receptors (out of 9) classified into the aminergic group (GPCRs 20, 21 and 25). As mentioned above, despite the deviations the algorithm assigned the aminergic binding site the lowest p-value.

All three methods, independently and in combination with 'Iterative signature algorithm', seek also for 'second order' results – results that explain different binding profiles within sub-groups of the GPCRs (and optionally – ligands). The second iteration of 'Correlated distances' method (the only method that succeeded in separating perfectly between aminergic and non-aminergic receptors) predicts that binding-specificity within non-aminergic receptors depends on four different residues, combined with eight amino-acid property clusters. Only one of the four (position

1304) is predicted by kPROT to face the interior of the protein. Position 1304 also has the lowest mean p-value (of the four results) and it is located in the part of the GPCR closer to the extracellular side, near the aminergic binding site. Therefore it can be treated as a strong candidate binding-site for non-aminergic GPCRs. No significant results were found for the aminergic group by any of the methods.

The lack of significant results may be taken to indicate that there are no specific position/s and amino-acid property/ies highly correlated with the binding profiles within any subgroup (either directly, as calculated by 'Correlated distances' method or mediated by ligand properties, as calculated by 'Hierarchical clustering' method and 'Correlated properties' method).

GPCR sequence similarity analysis [14] reveals a possible division of the aminergic GPCRs into five subgroups according to their natural ligand - muscarinic acetylcholine (GPCRs 1-5); dopamine (GPCRs 6 and 11); serotonin (GPCRs 7, 8, 10, 13, 16 and 17); histamine (GPCRs 9 and 15) and adrenergic receptors (GPCRs 12 and 14). Nevertheless, none of the clustering methods used (three clustering methods applied to two optional binding data sets) yielded a similar clustering.

'Correlated distances' method was applied to the subgroup of 17 aminergic receptors and an artificially created distances matrix, where the distance between GPCRs with the same natural ligand is defined as 0 and with different natural ligand – as 1. Unlike the results with the real distances matrix (that didn't include any significant position-property combinations), 24 different binding sites were identified using Bonferroni correction with p-value 0.05 (71 with FDR correction). Out of the 24 predicted binding-sites, 10 are predicted by kPROT to face the interior of the protein and the majority is located closer to the extracellular side.

These results indicate that the methods are able to identify binding sites, but are highly dependent on the input ligand binding activity data set. In the case of a general binding site such as the aminergic binding site, the data set must include both bio-amine ligands and non bio-anime ligands, as well as animergic GPCRs and non aminergic GPCRs. Since this requirement was fulfilled, the aminergic binding site was identified. However, the same condition applies for more specific binding sites, and apparently this requirement was not fulfilled in the used data set. The data set included at least two GPCRs of each kind, but probably doesn't include enough ligands of each kind to yield significant results.

Chapter 4
Summary and future work

This work suggests three optional methods for identifying binding-specificity determining residues, amino-acid properties and optionally ligand-properties (in 'Correlated properties' method only). In addition, we suggest an enhanced version of the methods using availabe clustering algorithms designed for expression data.

All methods are based on the assumption that if binding specificity is determined by an amino-acid property and a position, then the values of that property w.r.t the amino-acids in the position would be correlated with the ligand binding activity of the receptors.

The methods were applied to a varied dataset of 26 GPCRs, out of which 17 are bio-amines receptors. All methods were successful in identifying the known binding site of aminergic receptors, and some of them were also successful in identifying the amino-acid property related to the site - charge. Additional unknown binding sites and amino-acid properties were identified, both for the entire data set and for subgroups.

All the methods are general and can be used for any family of proteins, provided that they share a similar structure and sufficient data is available about them: namely measurements of their activity following the binding of a common set of ligands; their sequences and sequence alignment, and for some of the methods – a set of properties' values w.r.t the ligands.

Like most statistics-based procedures, the success of the suggested methods depends on the size of the data, and also on its diversity. Although the procedures succeeded in identifying the aminergic binding site with a small set of only 26 receptors, we expect to receive more results with higher accuracy given a larger data set.

Additional existing tools that predict binding sites or other related properties can be incorporated into the methods to increase their power. Structural properties such as the angular orientation of the selected position (as predicted by kPROT) and proximity to the extracellular side of the protein (as predicted by superimposing the positions upon the 3D structure of rhodopsin) can be taken into account in the positions' scoring (these properties were presented in this work for some of the resulting positions but not incorporated into the algorithms). Evolutionary properties can also aid in identifying binding sites. For example, ligand specificity-determining positions are expected to be conserved among orthologous receptors but not among paralogs (although this hypothesis is more strongly related to olfactory receptors, due to evolutionary pressure towards variedness, it is valid for other families of GPCRs as well to a lesser extent) [9]. A score positively correlated to the conservation among orthologous receptors and negatively correlated to conservation among paralogous receptors can be taken into account in the position scoring.

Finally, additional information regarding the original data can be used to "guide" the algorithm to the right results. For example, the 'Correlated properties' method matches amino-acid properties to ligand properties. Knowing the chemical meaning of the

given properties, we can eliminate pairs that are not expected to yield any correlation (such as amino-acid's charge and ligand's weight) and assign a higher weight to 'promising' pairs (such as amino-acid's charge and ligand's charge). Another example is using a pre-known classification of the receptors by enforcing it on the hierarchical clustering (either the clustering used by the 'Hierarchical clustering' method or the clustering induced by expression data clustering algorithms combined with any of the methods).

References

[1] Baldwin, J.M. The probable arrangement of the helices in G protein-coupled receptors. *EMBO J.* 12(4):1693-703, 1993.

[2] Brink CB, Harvey BH, Bodenstein J, Venter DP, Oliver DW. Recent advances in drug action and therapeutics: relevance of novel concepts in G-protein-coupled receptor and signal transduction pharmacology. *Br J Clin Pharmacol*. 57(4): 373-87. 2004.

[3] Branden C. and Tooze J., *Introduction to Protein Structure*. Garland Publishing, 1999. 251-281 pp.

[4] Alberts, B. et al. *The molecular biology of the cell*. Garland Publishing, 2002. 852-871 pp.

[5] Palczewski K. et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*. 2000 Aug 4; 289(5480):733-4.

[6] Strader C.D, Sigal I.S., Register R.B., Candelore M.R., Rands E and Dixon, R.A. Identification of residues required for ligand binding to the β-adrenergic receptor. *Proc. Natl. Acad. Sci. USA*, 84: 4384-4388, 1987.

[7] Sokal R.R. and Rohlf F.J. *Biometry*. W. H. Freeman and company, New York. 823-825 pp.

[8] Benjamini, Y. and Hochberg Y. Controlling the False Discovery Rate – a practical and powerful approach to multiple testing. *J. Roy Stat Soc B Met* 57(1): 289-300, 1995.

[9] Man, O., Gilad, Y and Lancet, D., Prediction of the odorant binding site of olfactory receptor proteins by human–mouse comparisons. *Protein Science*, 13:240–254, 2004.

[10] Bergmann S., Ihmels J and Barkai N., Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E67*, 031902, 2003.

[11] Pilpel, Y., Ben-Tal, N., Lancet, D., kPROT – A Knowledge-based Scale for the Propensity of Residue Orientation in Transmembrane Segments. Application to Membrane Protein Structure Prediction. *J. Mol. Biol*., 294:921-935, 1999.

[12] Shi L. and Javitch J.A., The binding site of aminergic G protein-coupled receptors: The Transmembrane segments and second extracellular loop. *Annu. Rev. Pharmacol. Toxicol*. 42:437-67, 2002.

[13] Getz, G., Levine E. and Domany, E., Coupled Two-Way Clustering Analysis of Gene Microarray Data. *PNAS 97*, 12079, 2000.

[14] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.* 215:403-410.