TEL AVIV UNIVERSITY אוניברסיטת תל-אביב
RAYMOND AND BEVERLY SACKLER FACULTY OF EXACT SCIENCES
SCHOOL OF PHYSICS AND ASTRONOMY
הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר
בית-הספר לפיזיקה ולאסטרונומיה

Unsupervised Extraction of Structures in Biological Data Sets

Dissertation submitted towards the degree of

Doctor of Philosophy

by

Assaf Gottlieb

Submitted to the Senate of Tel Aviv University

September 2009

This work was carried out under the supervision of

Professor David Horn

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. David Horn, who has guided me through both my M.Sc. and Ph.D. studies. His supervising approach allows for autonomous research and exploration but at the same time keeps careful supervision and guidance. This approach enabled me to develop independent thought, while at the same time know that his door is always open should I need assistance. I would like to thanks David also for his excellent remarks and the delegation of ideas, his ability to mark the diamonds out of clutter and his help in clear formulation of our concepts.

I would like to thank Judy and Stewart Colton for supporting me through the Colton Scholarship Fund. Without their support, this research would have been hard to conduct. Their motto of investing in people and not in buildings was proven indispensable in my case.

I would also like to thank collaborators and colleagues for helping me on my research projects. Dr. Roy Varshavsky pushed the UFF concept tremendously such that it might not have reached maturity without his efforts. Prof. Michal Linial and Prof. Nati Linial provided very helpful remarks and discussions regarding the UFF concept. Dr. Tsviya Olender enabled me quick entrance into the domain of Olfactory Receptors providing helpful remarks and sharing her vast knowledge of the subject. Last I would like to thank my colleague Uri Weingart for fun discussions regarding both academic and non-academic topics.

Last but not least, I would like to thank Idit, my dear wife and friend, who fully supported me on my decision to start my Ph.D. and been there for me the entire period. I would also like to thank my beloved children, Shachar and Noga, the former raised and the latter born throughout my Ph.D. studies and who are still giving me much joy in life.

Abstract

The amount and variety of data in natural sciences increases rapidly. Data abstraction, data manipulation and pattern discovery techniques are of great need in order to deal with such large quantities. Integration between different sources of data is also of major interest, as complex relations may arise. Biology is a good example of a field that provides extensive, highly variable and multi-sources data.

Extraction of patterns from data is often carried out in a supervised manner by matching data to prior knowledge (e.g. matching groups to known tags). Unsupervised pattern extraction, on the other hand, explores and identifies patterns inherent to the data, without additional prior knowledge. The vast amount of biological data, typically lacking extensive prior knowledge, makes it difficult to extract meaningful information. This fact provides the basis for unsupervised data exploration and pattern finding in biological data.

This thesis focuses on two topics that make use of unsupervised data analysis:

1. Unsupervised data mining algorithms and tools.
2. Analysis of protein families through unsupervised extraction of motifs.

The first topic includes methods for data exploration and pre-processing, typically referred to as data mining techniques. We present a novel dimensionality reduction framework termed unsupervised feature filtering (UFF). We apply UFF to various biological datasets, including cancer, HIV and Hepatitis-C gene-expression datasets and cancer microRNA expression arrays. Using the UFF selected features for clustering enable us to reduce noise and achieve clear clusters, which match known instance tagging, when this information is available. Furthermore, the selected sets of genes and microRNAs show enrichment of both related and surprising terms. Most of the top ranked genes and microRNAs have documented relations to the specified disease while for others, these relations are yet undetermined. These selected sets may thus contain true biological meaning.

The second topic deals with deterministic sequence motifs, extracted by the Motif Extraction (MEX) algorithm. We develop a method to construct a meaningful set of these deterministic motifs termed Common Peptides (CPs). This set forms a framework, enabling exploration of various protein families, revealing internal protein family clusters, finding historical traces of evolutionary events and exposing remote homology between proteins. This framework was applied to Olfactory Receptors (ORs) and to the enzyme families of aminoacyl-tRNA synthetases (aaRS). Using the CP framework on ORs we track OR evolutionary events in vertebrates, revealing redundancy removal in humans relative to other mammals, the mass losses in the reptiles lineage and the history of OR

families. We also point out CPs that differentiate between water and land dwelling species and identify their specific locations on the OR sequence.

Using the CP framework on aaRS families reveal different distribution of aaRS families across the different kingdoms of life. This framework also identifies CPs that differentiate between the two known classes of the aaRS families, including many unnoticed sequence motifs. Abundant CPs tend to overlap known catalytic and binding regions.

Contents

Chapter 1

General Introduction

1.1 Introduction

In many disciplines, data comes in many flavors and shapes. The rapid increase of available data in biology requires the development of techniques to control the data, to separate the wheat from the chaff and to arrange it in a way that is presentable.

As the data grows more complex, possibly containing inherent noise and irrelevant features, selecting the best techniques suitable for the problem, tailoring them together and modifying them to answer the problem at hand are crucial.

The techniques subjected to the general term of data exploration are traditionally separated into groups, such as supervised and unsupervised learning, feature selection and extraction and pattern extraction.

While supervised learning has been studied extensively, typically borrowed from other disciplines to study biological datasets, unsupervised learning also plays an important, yet less studied, role in the processing and exploration of the data. As unsupervised learning is primarily concerned with the data itself, different solutions are often tailored to a specific data type or even to a specific data-set.

In the past years, automation of biological data extraction has rapidly increased, introducing vast amounts of un-annotated data-sets. One example of such biological data-sets is expression microarrays, measuring expression of genes, microRNAs or proteins in a certain cellular environment. Another example is DNA and protein sequences of multiple species. This thesis confronts primarily these two aforementioned biological data types and develops novel unsupervised solutions that enable extracting meaningful patterns from them.

1.2 Thesis outline

This thesis begins with Chapter 1, a general introduction, providing a brief survey of the main tasks this thesis deals with.

Following the introduction, this thesis is divided into two distinct parts. Part one is dedicated to feature selection. It includes a short introduction to feature selection and dimensionality reduction (chapter 2), followed by the presentation of the novel Unsupervised Feature Filtering (UFF) algorithm in chapter 3. UFF takes into account the interplay between different features by ranking them according to the influence of each feature on a global function calculated over all other features. In chapter 4 we analyze UFF selected features and describe a framework encompassing UFF. This framework provides measures to assess the quality of the UFF selected features, enhances its performance and implements the entire framework as a web tool.

Part two of this thesis introduces the concept of Common Peptides (CPs) – a semi-supervised method that exploits the unsupervised Motif Extraction (MEX) algorithm to produce sets of deterministic motifs from protein families. It is described in chapter 5. Chapter 6 introduces a specific application of the CP methodology to produce interesting insights of vertebrate Olfactory Receptors (ORs). Chapter 7 applies the same CP framework to a family of enzymes called aminoacyl tRNA synthetases, an important building block of the DNA translation to proteins mechanism.

The final chapter concludes this thesis and provides a summary of the presented algorithms and methods and some further insights.

Chapters 3, 4, 6 and 7 are based on published or submitted manuscripts. All of them are presented as separate units, containing their own references, figures and tables to enhance readability.

Part 1

Chapter 2

Introduction to feature selection

2.1 Introduction

An important aspect of data analysis includes dimensionality reduction of the data. This can be viewed as a preprocessing task preceding the data analysis or even as a significant part of the data analysis itself, providing valuable insight regarding underlying patterns in the data. According to [1-3], dimensionality reduction objectives are to improve model performance, reduce over-fitting and lower running time and other resources. The introduction of high-throughput technologies produces huge-sized datasets, where dimensionality reduction is crucial.

It is customary to divide dimensionality reduction methods to *feature extraction*, where the methods transform all, or a part of the features to a lower dimension space. Conversely, *feature selection* methods select a subset of the original features.

In many disciplines and in Biology in particular, feature selection methods bear a significant advantage over feature extraction methods. This advantage is the capability to attach meaning to the selected features, connecting them to the relevant analysis of the data. In biological data-set analysis, these features may be defined as testable biomarkers, reducing the cost of testing the entire set of features for each new sample (e.g. a set of genes for a new patient).

Most of the existing methods of feature selection are supervised, i.e. selecting features that match a predefined labeling of the samples. Unsupervised feature selection methods are few [3, 4]. In an analogous way to the supervised methods, unsupervised methods also divide to 3 types, according to where they take place: before, during or after the clustering procedure of the samples. The methods occurring before the clustering are called *filtering* methods.

Feature filtering methods are considered to be the least biased of the three, being independent of subsequent data analysis procedures such as the type of clustering algorithm. Most of the unsupervised feature-filtering methods operate on a single feature at a time, calculating some function on the feature values for all training samples (e.g. feature variance, maximum to minimum ratio (fold) or entropy), ignoring the interplay between features.

Chapter 3 introduces a novel Unsupervised Feature Filtering (UFF) method, which scores features based on relation to all other features in the dataset. Furthermore, it provides a natural cutoff to decide how many features to choose. Chapter 4 extends UFF by examining the type of features it selects and provides a framework which enables the implementation of UFF as a web-tool.

2.2 References

1. Guyon I, Elisseeff A: **An Introduction to Variable and Feature Selection**. *Journal of Machine Learning Research* 2003, **3**:1157--1182.
2. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.
3. Liu H, Li J, Wong L: **A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns**. *Genome Inform* 2002, **13**:51-60.
4. Dy JG, Brodley CE: **Feature Selection for Unsupervised Learning**. *J Mach Learn Res* 2004, **5**:845-889.

Chapter 3

Unsupervised Feature Filtering (UFF)[1]

3.1 Introduction

Feature selection is an important tool in many biological studies. Given the large complexity of biological data, e.g. the number of genes in a microarray experiment, one naturally looks for a small subset of features (e.g. small number of genes) that may explain the properties of the data that are being investigated. This type of motivation fits into the general scheme of **feature exploration**, i.e. searching for features because of their direct biological relevance to the problem. An alternative motivation is that of **preprocessing**: searching for a small set of features to simplify computational constraints, to allow for the handling of high throughput biological experiments, and to separate signal from noise. Practically, selection of a small set of genes is of ultimate importance when a small set of informative genes can be the basis for cancer diagnosis and a basis for development of gene associated therapy.

Preprocessing often involves some operation on feature-space in order to reduce the dimensionality of the data. This is referred to as **feature extraction,** e.g. restricting oneself to the first *r* principal components of a PCA routine. Note that superpositions of features appear in this example. Alternatively, in **feature selection** we limit ourselves to particular features of the original problem. This is the subject to be studied here. Let us refer to [1] for a comprehensive survey.

It is conventional to distinguish between **wrapper** and **filter** modes of the feature selection process. Wrapper methods contain a well-specified objective function, which should be optimized through the selection. The algorithmic process usually involves several iterations until a target or convergence is achieved. **Feature filtering** is a process of selecting features without referring back to the data classification or any other target function. Hence we find filtering as a more suitable process that may be applied in an **unsupervised** manner.

Unsupervised feature selection algorithms belong to the field of unsupervised learning. These algorithms are quite different from the major bulk of feature selection studies that are based on supervised methods (e.g., [1, 2], and compared to the latter are relatively overlooked. Unsupervised studies, unaided by objective functions, may be more difficult to carry out, nevertheless they convey several important theoretical advantages: they are unbiased, by neither the experimental expert nor by the data-analyst, can be preformed well when no prior knowledge is available, and they reduce

[1] Based on the paper *Novel Unsupervised Feature Filtering of Biological Data,* Roy Varshavsky, Assaf Gottlieb, Michal Linial and David Horn, Bioinformatics 2006, 22(14):e507-513 (Presented in ISMB 2006).

the risk of overfitting (in contrast to supervised feature selection that may be unable to deal with a new class of data). The downside of the unsupervised approach is that it relies on some mathematical principle, like the one to be suggested in this study, and no guarantee is given that this principle is universally valid for all data. A common practice to resolve this quandary is to demonstrate the success of the method on various biological datasets and compare the results obtained by the method with external knowledge.

Existing methods of unsupervised feature filtering include ranking of features according to range or variance (e.g., [3], [1], selection according to highest rank of the first principal component ('Gene shaving' of [4, 5] and other statistical criteria. An example of the latter is [6] where all possible partitions of the data are considered and the corresponding features are labeled. The partitions with statistical significant overabundance are selected. Another example is of [7], who optimize a function based on the spectral properties of the Laplacian of the features.

Here we present an intuitive, efficient and deterministic principle, leaning on authentic properties of the data, which serves as a reliable criterion for feature ranking. We demonstrate that this principle can be turned into efficient and successful feature selection methods. They compete favorably with other popular methods.

3.2 Methods

3.2.1 Mathematical framework and notations

Let us consider a dataset of n instances[2] $A_{[nXm]} = \{\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_i, \ldots, \bar{A}_n\}$, where each instance, or observation, $\bar{A}_i$ is a vector of m measurements or features. The objective is to define a subset of features $\tilde{M}$, of size $m_c<m$, that, in a sense to be defined below, best represents the data.

In PCA (or SVD) studies it is conventional to regard the best representation as the minimal least-square approximation of the original matrix [8]. This principle can be followed also in feature extraction but it has the disadvantage that it may preserve too many properties of the data, including systematic noise. We will define our 'best approximation' using a principle based on SVD-entropy, and subject it to an a-posteriori test: given different selection rules of features choose the ones that prove useful as basis for the best fit to labeled data, e.g., perform clustering within the data-space spanned by the selected features and compare the results with known classification. This comparison will be performed using the Jaccard score.

$$J = \frac{n_{11}}{n_{11} + n_{01} + n_{10}} \quad (1)$$

[2] In this paper A (or $A_{[nXm]}$) is a matrix and $\bar{A}$ (or $\bar{A}_i$) is a vector.

where n_{11} is the number of pairs of instances that are classified together, both in the 'expert' classification and in the classification obtained by the algorithm; n_{10} is the number of pairs that are classified together in the 'expert' classification, but not in the algorithm's classification; n_{01} is the number of pairs that are classified together in the algorithm's classification, but not in the 'expert' classification;

The Jaccard score reflects the 'intersection over union' between the algorithm's clustering assignments and the expected classification. Its values range from 0 (no match) to 1 (perfect match).

3.2.2 Ranking by SVD-Entropy

[9] have defined an SVD-based entropy of the dataset. Denote by s_j the singular values of the matrix A. s_j^2 are then the eigenvalues of the nxn matrix AA^t. Let us define the normalized relative values [8]:

$$V_j = s_j^2 / \sum_k s_k^2 \qquad (2)$$

and the resulting dataset entropy [9]:

$$E = -\frac{1}{\log(N)} \sum_{j=1}^{N} V_j \log(V_j) \qquad (3)$$

This entropy varies between 0 and 1. E = 0 corresponds to an ultra-ordered dataset that can be explained by a single eigenvector (problem of rank 1), and E = 1 stands for a disordered matrix in which the spectrum is uniformly distributed.

Figure 1 demonstrates two examples of 5 eigenvalues, one with high entropy (left, 0.87) and the other with low entropy (right, 0.14). As can be seen in figure 1, when the entropy is very low, one expects a very non-uniform behavior of eigenvalues. One should not confuse the standard definition of entropy, based on probabilities [10], with the one used here, which is based on the distribution of eigen- (or singular) values. Although standard entropy considerations appear in feature selection methods, such as the supervised bottleneck approach [11], the use of SVD-entropy for feature selection is a novel approach.

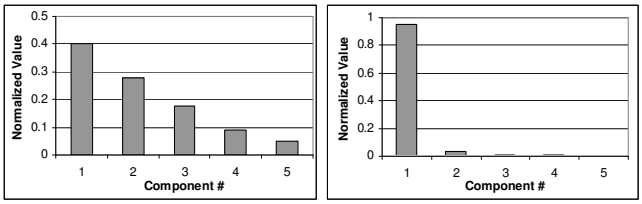


Figure 1: A comparison of two eigenvalue distributions; the left has high entropy (0.87) and the right one has low entropy (0.14)

We define the contribution of the *i-th* feature to the entropy (CE_i) by a leave-one-out comparison according to

$$CE_i = E(A_{[nXm]}) - E(A_{[nX(m-1)]}) \qquad (4)$$

where, in the last matrix, the *i-th* feature was removed.

Thus we can sort features by their relative contribution to the entropy. Let us define the average of all *CE* to be *c* and their standard deviation to be *d*. We distinguish then between three groups of features:

1. $CE_i > c+d$, features with high contribution
2. $c+d > CE_i > c-d$ features with average contribution
3. $CE_i < c-d$ features with low (usually negative) contribution

Features in the first group (high CE) lead to entropy increase; hence they are assumed to be very relevant to our problem. Retaining these features we expect the instances to be more evenly spread in the truncated SVD space. The features of the second group are neutral. Their presence or absence does not change the entropy of the dataset and hence they can be filtered out without much information loss. The third group includes features that reduce the total SVD-entropy (usually $c-d < 0$). Such features may be expected to contribute uniformly to the different instances, and may just as well be filtered out from the analysis.

The first feature selection method that we propose is to limit oneself to the first group of features according to the *CE* ranking. *A* will then be represented by a new matrix of rank m_c, the number of features in group 1. Several other feature selection methods are suggested in the next section. In all of them we assume that the same value of m_c continues to serve as the right guide for optimal dimensionality reduction.

3.2.3 Three Feature Selection Methods

Entropy maximization can be implemented in three different ways, as is also the case in other feature selection methods.

Simple ranking (SR): select m_c features according to the highest ranking order of their CE values.

Forward Selection (FS): here we consider two implementations.

FS1: Choose the first feature according to the highest CE. Choose among all other features the one which, together with the first feature, produces a 2-feature set with highest entropy. Continue with iteration over all *m-2* features to choose the third according to maximal entropy, etc, until m_c features are selected (Box 1).

FS2: Choose the first feature as before. Recalculate the *CE* values of the remaining set of size *m-1* and select the second feature according to the highest *CE* value. Continue the same way until m_c features are selected (Box 2).

Backward Elimination (BE): Eliminate the feature with the lowest CE value. Recalculate the CE values and iteratively eliminate the lowest one until mc features remain (Box 3).

```
1. Start with M̃ = ∅ and M′ = M
2. Select the element with the highest
   CE. Remove it from M′, insert it into M̃
3. While size of M̃ < m_c
  a. For each element in M′(∀m ∉ M̃) compute
     its CE score on M̃ (E(A_{M̃+i})–E(A_M̃))
  b. Select the element with the highest CE
     Score → remove from M′, insert into M̃
4. End
```

```
1. Start with M̃ = M and M′ = ∅
2. While size of M̃ > m_c
  a. Select the element in M̃ with the lowest
     CE Score
  b. Remove from M̃, insert into M′
3. End
```

Box 1: Pseudo-code of Forward Selection method FS1
Box 2: Pseudo-code of Forward Selection method FS2
Box 3: Pseudo-code of Backward Elimination method BE

One may view the different methods also as specifying alternative ranking methods. Whereas SR ranks the features according to their original CE values, FS1, FS2 and BE introduce other ranking orders through the algorithms defined above. In the examples studied below we display rankings for the entire range of 1 to *m.*

```
1. Start with M̃ = ∅ and M′ = M
2. While size of M̃ < m_c
  a. Select the element in M′(∀m ∉ M̃) with
     the highest CE Score
  b. Remove from M′, insert into M̃
3. End
```

In an appendix we analyze the computational complexity of all these methods. SR is the fastest one and BE is the most cumbersome one for large numbers of features. In the examples to be discussed next, we will compare the different methods with one another. However, because of complexity, the BE method will be used in only one of the examples.

3.3 Results

Our four feature filtering methods were compared with each other and with two known methods: Variance Selection (VS) and Gene Shaving (GS). The latter is a variation of a method of [4] which removes features iteratively according to their lowest correlations with the first principal component. For comparison we also look at results of random feature selection on several benchmarks.

3.3.1 The viruses dataset of Fauquet, 1988

This is a dataset of 61 rod-shaped viruses affecting various crops (tobacco, tomato, cucumber and others) originally described by [12] and analyzed more thoroughly by [13]. There are 18 measurements of Amino Acid Compositions (AAC) for the coat proteins of the virus that serve as 18 features. The viruses are known to be classified into four classes: Hordeviruses (3), Tobraviruses (6), Tobamoviruses (39) and Furoviruses (13). Figure 2 displays the CE values of all 18 features. Our criterion sets $m_c = 3$. We test the performance of the system for the entire m range to see if this choice makes sense. Before doing so, let us display the ranking orders of all methods in Table 1. By definition, SR has the same ranking order as CE in Figure 2. In this problem, BE turns out to lead to the same order as FS1, and all our three methods agree with each other on the first three features to be selected. We include in Table 1 also the ranking order of VS (variance selection) and GS (gene shaving). The two last ones are highly correlated with each other (Spearman correlation 0.76) but highly uncorrelated with our three methods (see the Supplementary Material section for more details). In particular note that VS chooses ASX and GLX as its second and third features, whereas for our three methods these two features are unfavorable (15^{th} to 18^{th}) choices.

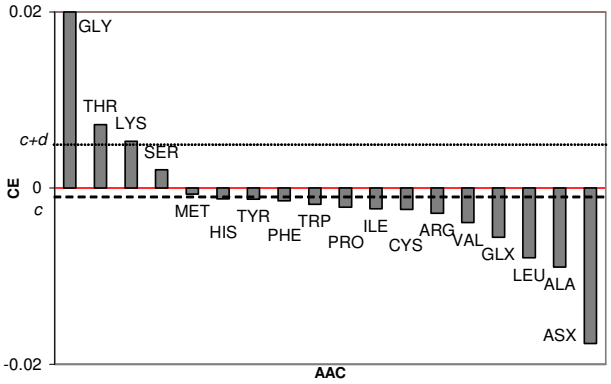


Figure 2: CE of the 18 Amino Acid Compositions (AAC) of the virus dataset. ASX stands for ASN and ASP and GLX for GLN and GLU. The dashed line represents the value of *c* and the dot-dashed line the value of *c+d*.

AAC	SR	FS1/BE	FS2	VS	GS
GLY	1	1	1	1	9
THR	2	2	2	6	6
LYS	3	3	3	4	14
SER	4	13	4	5	4
MET	5	4	15	16	17
HIS	6	6	7	15	16
TYR	7	8	13	13	13
PHE	8	7	5	14	11
TRP	9	5	16	17	15
PRO	10	11	6	11	10
ILE	11	10	11	12	12
CYS	12	9	18	18	18
ARG	13	12	10	8	8
VAL	14	14	8	9	7
GLX	15	16	9	3	2
LEU	16	15	14	10	5
ALA	17	17	12	7	3
ASX	18	18	17	2	1

Table 1: Ranking of the 18 Amino Acid Compositions of the virus dataset according to various feature filtering methods. Colors from white to black match the numbers that reflect the ranking of each method.

Next we evaluate the subset selection using the Jaccard score. This is done by applying the QC clustering algorithm [14] on the 61 viruses described by the selected subset of features. QC was applied after reduction of each space to normalized 3-space dimensions, using the parameter σ=0.5 (for details see [15], and COMPACT[3]). Results are shown in Figure 3 for three of our four methods. All three do exceedingly well at the three features level (J>0.9) whereas the variance method obtains J=0.4. Note that our methods, with our choice of mc, lead to a much better result than J=0.6,

[3] http://adios.tau.ac.il/compact or http://www.protonet.cs.huji.ac.il/compact

obtained when all 18 features are taken into account. This exemplifies the importance of keeping features that maximize the entropy. The feature ranking of FS1 and BE is the only one that keeps performing very well with more than three selected features. Similar relative successes of feature selection evaluation (although less favorable J-scores) were obtained with other clustering methods, such as K-means. This comparison, as well as other details that could not be fitted into this paper, can be found in the Supplementary Material[4].

[12] have argued that the AAC of the coat protein of plant viruses are specific to the structure of the viral particle, to the mode of transmission and to sub-grouping of viruses to distinctive classes. Our results indicate that choosing only 3-4 features correctly, not only preserves the classification but allows much better performance with minimal failure. It is interesting to note that the 3 highest-ranking amino acids, GLY, THR and LYS are not dominating the coat proteins. These amino acids account for only 13-21.5% of the coat proteins, a fraction that is similar to the average percentage in the entire proteins database (18.3%). Further investigation shows that neither their size nor polarity or electric charges differentiate these three amino acids from the remaining. Nevertheless, since GLY, THR, LYS and MET (the fourth ranked AAC, according to the FS1 method) represent different functional groups, we conclude that the FS1/BE ranking is consistent with selecting amino acids that carry different physico-chemical properties.

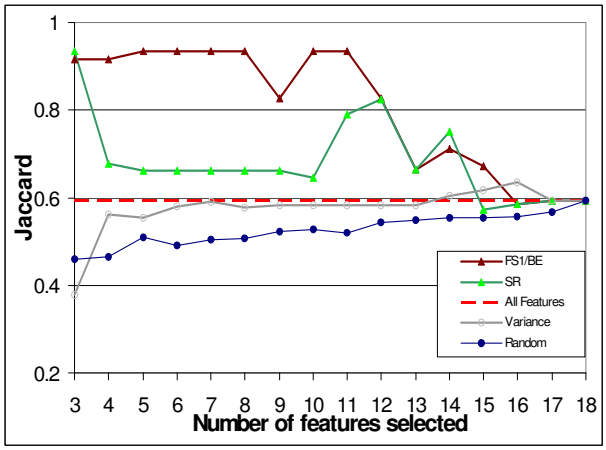


Figure 3: Filtering quality of the virus dataset is tested by Jaccard scores of clustering performed in spaces spanned by them (see text). Best results are obtained for FS1 (identical with BE in this case) and SR for m_c=3. FS1 continues to perform very well with more features. Feature selection according to VS performs worse. For comparison we include also an evaluation based on a large group of random order rankings.

[4] http://adios.tau.ac.il/compact/UFF/SUPP

3.3.2 The MLL dataset of Armstrong et al., 2002

The second dataset that we apply our methods to is that of Armstrong et al., 2002, who have attempted to cluster data of three Leukemia classes: lymphoblastic Leukemia with MLL translocations and conventional acute lymphoblastic (ALL) and acute myelogenous Leukemias (AML). In the experiment, 12582 gene expressions were recorded, using Affymetrix U95A chips on 72 patients, 20 of which diagnosed as MLL, 24 ALL and 28 AML. They showed that these 3 Leukemia types can be divided according to some gene expression. However, when filtering in an unsupervised manner (selecting 8700 genes that show some variability in expression level), the clustering results were unsatisfactory and much inferior to a supervised selection of 500 genes that best separate between the cancer patients.

Applying our CE criteria we use the method SR, and compare clustering of these feature-filtered data with VS (Figure 4). Clustering was performed by K-Means, averaging over 100 runs and using K=3 with data projected onto a unit sphere in 3D-reduced space [15]. The asymptotic Jaccard score is J=0.426 for this K-Means method. As can be seen in Figure 4 VS provides no improved quality, whereas SR leads to J-values between 0.7 and 0.8 for filtered gene groups of sizes 250 to 450. The preferred m_c value according to $c+d$ of SR is 254. Better results can be obtained by using the QC algorithm, but the same trend and conclusions regarding feature selection hold also there. It is interesting to note that QC clustering of our unsupervised SR method, for m_c=254, reaches J=0.85 (see supplementary).

We display the K-Means analysis in Figure 4, in spite of its poorer performance compared to QC, in order to emphasize that the quality of the feature filtering method is independent of the clustering-test performed on the filtered data.

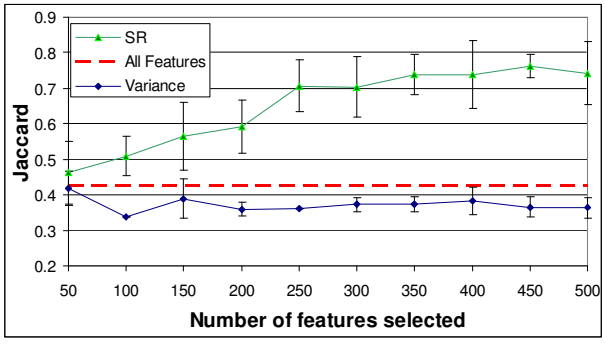


Figure 4: Clustering quality of two feature selection methods. Results are averages of 100 runs of K-Means clustering.

3.3.3 The Leukemia dataset of Golub et al., 1999

After demonstrating the effectiveness of our methods on both small and large datasets, we choose a third dataset [16] that has served as a benchmark for several clustering algorithms ([17, 18] and more} and feature selection methods (e.g., [2, 19]. The experiment sampled 72 Leukemia patients with two types of Leukemia, ALL and AML. The ALL set is further divided into T-cell Leukemia and B-cell Leukemia and the AML set is divided into patients who have undergone treatment and those who did not. For each patient, an Affymetrix GeneChip measured the expression of 7129 genes. The task is clustering into the four correct groups within the 72 patients in a [7129x72] gene-expression matrix. This clustering task is quite difficult. Using the QC method (in normalized 5 dimensions with σ=0.54), applied to the data without feature selection, one obtains J=0.707, which is the best score for a variety of clustering algorithms [15].

The CE values for the 7129 features of this problem are displayed in Figure 5. Most of the features have a zero score. There are about 150 large CE values (see Figure 5) and about the same number of small CE values. The bright color within the inset indicates the first 100 features selected by FS1. While their ordering is different from the SR ranking, most of them belong, as expected, to the class of large CE values. The overlaps of the first leading features of SR with those of FS1 and FS2 are shown in the Venn diagrams of Figure 6.

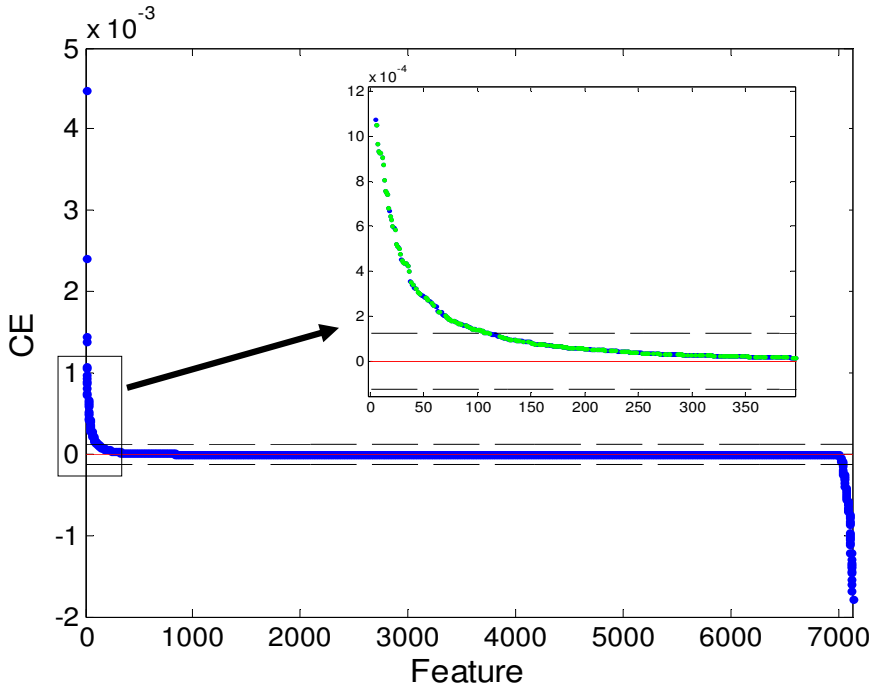


Figure 5: CE of the 7129 genes of the Golub dataset (c=0, dashed lines represent $c \pm d$). The inset zooms into the highest-ranked 300 genes, with bright dots signifying the top 100 features according to the FS1 method.

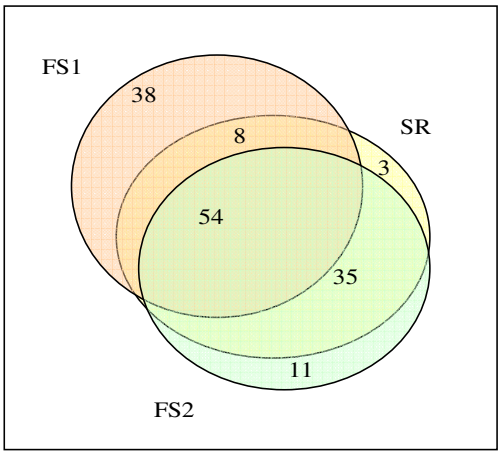


Figure 6 : Venn diagram of relations among the first 100 features selected by different methods.

Next we turn to testing the filtering methods to see how well they do in the clustering task, i.e. what are the Jaccard scores that are obtained by applying an identical clustering algorithm to the different spaces spanned by the selected features. The clustering algorithm is the QC method mentioned above. Figure 7 shows that good results can be obtained by our filtering methods once the gene subset is larger than 100 or so. For feature sets of sizes 120 to 200 we find selections (of FS1 and SR) that lead to Jaccard scores that are better than J=0.707, the asymptotic limit. Gene subsets larger than 300 result in Jaccard scores below the asymptotic limit (for a complete list, see the supplementary material). Also in this problem the GS results are inferior to those of the other methods.

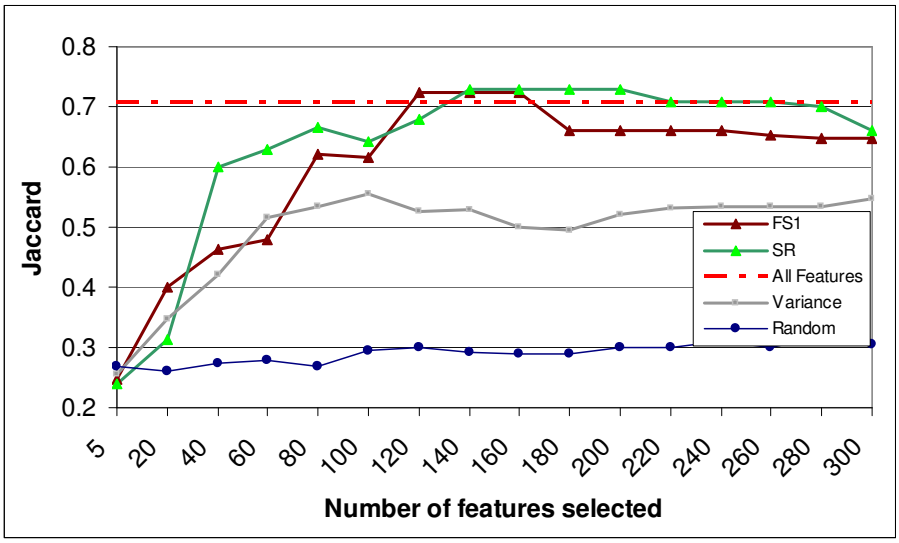


Figure 7 :Jaccard scores of QC clustering for different feature filtering methods on small gene subsets of the Golub data.

3.3.3.1 Biological interpretations of the Leukemia dataset of Golub et al., 1999

It is clearly of interest to look at the 100 or so genes that participate in the sections that lead to the best Jaccard score. In Figure 6 we saw that there exists a substantial overlap between the choices of our three different methods. To study the biological significance of our subset of overlapping 54 genes we have run a GO enrichment analysis (NetAffx™ web tool[5]) on this subset. As displayed in Figure 8 (and supplementary), we are able to assign some prevalent biological processes to the selected genes.

The association of our selected 54 genes with functional annotation related to defense, inflammation and response to pathogen (with p-value ranging for e^{-7} to e^{-22}) is intriguing (Figure 8). It may underlie the difference in AML and ALL in view of the different susceptibility of the patients to treatment such as chemo and radiotherapy. Thus the listed protein processes may not only be considered as 'subtype cancer markers' but as an indication of the biological properties of the cancerous cells. Specifically, cellular response to pathogen, to stress and to inflammation may be different for AML and ALL. It may also provide a focused hypothesis towards the processes and mechanisms that can be used as a follow up in monitoring the outcome of therapy in case of Lymphoma.

[5] http://www.affymetrix.com/analysis/index.affx

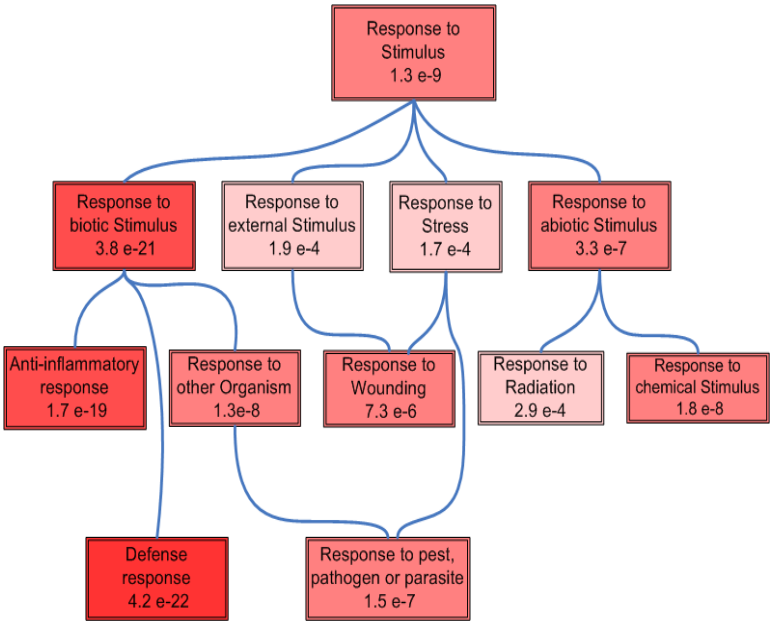


Figure 8: Diacyclic graph of GO enrichment. Shown are GO nodes [20] with significant p-value of enrichment as determined by the NetAffx™ tool[5] (p-value < 5e-4). The color of each node matches its significance level (along the spectrum of red shades, light: lowest to dark: highest).

3.4 Discussion

We have introduced a novel principle for unsupervised feature filtering that is based on maximization of SVD-entropy. The features can be ranked according to their CE-values. We have proposed four methods based on this principle and have tested their usefulness on three different biological benchmarks. Our methods outperform other conventional unsupervised filtering methods. This is clearly brought out by the examples that we have analyzed. More details are provided by our Supplementary Material[6]. In particular, it is striking to note how much more successful our methods are compared to VS, the popular variance ordered method.

The major theoretical difference between the two approaches is that VS relies on a measurement of one feature at a time. The entropy-based approach, as implemented by the CE calculation, takes into account the interplay of all features. In other words, the contribution of a feature, its CE, depends on the behavior of all other features in the problem. Thus variance is only one of the factors that affect the CE value. The CE value depends also on the correlations (or the absence thereof) of a given

[6] http://adios.tau.ac.il/compact/UFF/SUPP

feature with all others. The difference between the ranking of SR and VS in Table 1 bears evidence to the difference between the two methods.

We have demonstrated that our selected features have important biological significance, through a GO enrichment analysis of the genes in the Golub dataset. A similar analysis of the Armstrong dataset is presented in the Supplementary Material[6]. In the virus dataset, we have shown that the FS1/BE filtering method works exceedingly well for a large range of numbers of features. The biological significance of the relevant choices of amino-acids remains to be uncovered.

The CE ranking leads to an estimate of the optimal m_c choice. This is an important point by itself. In other methods, such as VS, it is almost impossible to make this choice on the basis of variation of feature properties. Conventionally one makes therefore an arbitrary choice, such as selecting 10% or 50% of the features. In the three datasets discussed in our paper it seems quite clear that our suggested optimal m_c, as judged from the CE scores, leads indeed to optimal results. The improved Jaccard scores indicate that the selected m_c features have biological significance.

Our four methods differ in computational complexity. SR is the simplest one, since it relies just on sorting the initial CE values. In an appendix we compare its complexity with that of the other methods. The relative values depend on the choice of m_c (the size of the subset).

FS1 chooses features that lie high on the original CE-score, hence its optimal selected set will have a large intersection with that of SR. Nonetheless, for small numbers of selected features, the order may be very important. Thus, in the virus problem, FS1 turns out to be much more successful than SR. In the Leukemia datasets, where reasonable results were obtained for larger feature sets, FS1 was not found to be significantly better than SR. Biologically one may expect the appearance of features that are degenerate with one another, i.e. have quite identical behavior on all instances. Such duplicity can be included by the SR method but excluded by the FS1 one.

Our optimal feature-filtered sets in the two Leukemia problems turn out to include just few percents of all genes. Thus a CE-analysis indicates that a small subgroup of all genes is the most relevant one to the data in question. We have seen that this relevance is borne out by both Jaccard scores and GO enrichment analysis. The pursuit of small feature sets is often guided by wishful thinking that the essence of biological importance can be reduced to a small causal set. Here we find that the small number obtained in our analysis is an emerging phenomenon, and may be regarded as a true biological result.

3.5 References

1. Guyon I, Elisseeff A: **An Introduction to Variable and Feature Selection**. *Journal of Machine Learning Research* 2003, **3**:1157--1182.

2. Liu H, Li J, Wong L: **A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns**. *Genome Inform* 2002, **13**:51-60.
3. Herrero J, Diaz-Uriarte R, Dopazo J: **Gene expression data preprocessing**. *Bioinformatics* 2003, **19**(5):655-656.
4. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P: **'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns**. *Genome Biol* 2000, **1**(2).
5. Ding C, He X, Zha H, Simon H: **Adaptive dimension reduction for clustering high dimensional data**. *IEEE International Conference on Data Mining* 2002:107-114.
6. Ben-Dor A, Friedman N, Yakhini Z: **Class Discovery in Gene Expression Data**. *RECOMB* 2001.
7. Wolf L, Shashua A: **Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weight-Based Approach**. *JMLR* 2005, **6**:1855–1887.
8. Wall M, Rechtsteiner A, Rocha L: **Singular Value Decomposition and Principal Component Analysis**. In: *A Practical Approach to Microarray Data Analysis*. Edited by Berrar D, Dubitzky W, Granzow M: Kluwer; 2003: 91-109.
9. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *PNAS* 2000, **97**(18):10101-10106.
10. Shannon C: **A mathematical theory of communication**. *The Bell system technical journal* 1948, **27**:379-423, 623-656.
11. Tishby N, Pereira FC, Bialek W: **The Information Bottleneck Method**. *Proc of the 37-th Annual Allerton Conference on Communication, Control and Computing* 1999:368-377.
12. Fauquet C, Desbois D, Fargette D, Vidal G: **Classification of furoviruses based on the amino acid composition of their coat proteins**. *Dev Appl Biol* 1988, **2**:19-36.
13. Ripley BD: **Pattern Recognition and Neural Networks**. Cambridge: Cambridge University Press; 1996.
14. Horn D, Gottlieb A: **Algorithm for data clustering in pattern recognition problems based on quantum mechanics**. *Phys Rev Lett* 2002, **88**(1):018702-018702.
15. Varshavsky R, Linial M, Horn D: **COMPACT: A Comparative Package for Clustering Assessment**. In: *Lecture Notes in Computer Science*. 3759 edn: Springer-Verlag; 2005: 159-167.
16. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al*: **Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**. *Science* 1999, **286**(5439):531-537.
17. Sharan R, Shamir R: **CLICK: a clustering algorithm with applications to gene expression analysis**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:307-316.
18. Getz G, Levine E, Domany E: **Coupled two-way clustering analysis of gene microarray data**. *Proc Natl Acad Sci U S A* 2000, **97**(22):12079-12084.
19. Liu B, Cui Q, Jiang T, Ma S: **A combinational feature selection and ensemble neural network method for classification of gene expression data**. *BMC Bioinformatics* 2004, **5**:136.
20. Camon E, Barrell D, Lee V, Dimmer E, Apweiler R: **The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase**. *In Silico Biol* 2004, **4**(1):5-6.
21. Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, Croz JD, Greenbaum A, Hammarling S, McKenney A *et al*: **LAPACK User's Guide** (**http://www.netlib.org/lapack/lug/lapack_lug.html)**, Third edn. Philadelphia: SIAM; 1999.

3.6 Appendix

3.6.1 Computational complexity of the four methods

In the following calculations, we will assume that $m_c<n$, which will give upper bound to the complexity. We will not assume that $m<n$.

The computation of all eigenvalues for a dense symmetric matrix requires $O(p^3)$ operations, where p is the size of the matrix [21].

We will define the complexity of the initial computation of all CEs to be $O(m*min(n,m)^3) \equiv K$.

SR: The computational complexity is lowest for the SR method. There's only one calculation of all CEs, followed by sorting. Hence the complexity is $O(K + m*logm)$.

FS1: Calculation of all CEs followed by $(m_c - 1)$ repetitive diagonalization of a growing matrix (from 2 to $(m_c - 1)$), leading to $O(K + m*m_c^4)$.

FS2: Calculation of all CEs followed by $(m_c - 1)$ repetitive diagonalization of a decreasing matrix (from $m-2$ to $(m-m_c)$), leading to $O(m^5-(m-m_c)^5)$. Note that here, if $n< (m-m_c)$, the complexity is $O(mm_cn^3)$

BE: Calculation of all CEs followed by $(m-m_c-1)$ repetitive diagonalization of a decreasing matrix (from $m-2$ to (m_c-1)), leading to $O(m^5-m_c^5)$. Note that here, if n<m, the complexity is reduced to $O((m^2-m_c^2)n^3)$.

Clearly computational complexity is lowest for the SR method, since only one calculation of all CEs is needed. BE or FS2 have the highest complexity, depending on whether $m>2m_c$ or not.

3.7 Supplementary Material

Figures S1-S13 and GO enrichment table are also found in http://adios.tau.ac.il/compact/UFF/SUPP and in the additional CD.

3.7.1 The Viruses dataset of Fauquet, 1998

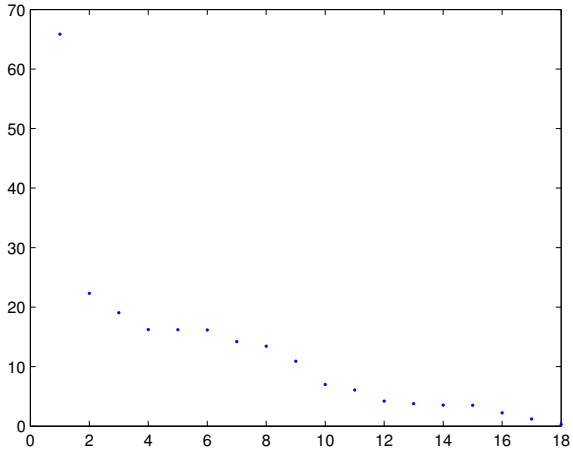


Figure S9: Variance of the features of the virus dataset

Spearman	SR	FS1	FS2	BE	VS	GS
SR	1	0.8824	0.63	0.8824	-0.0114	-0.4572
FS1	0.8824	1	0.4056	1	-0.2384	-0.676
FS2	0.63	0.4056	1	0.4056	0.4861	0.162
BE	0.8824	1	0.4056	1	-0.2384	-0.676
VS	-0.0114	-0.2384	0.4861	-0.2384	1	0.7647
GS	-0.4572	-0.676	0.162	-0.676	0.7647	1

Table S2: Spearman correlation of the features ranking according to the various selection methods

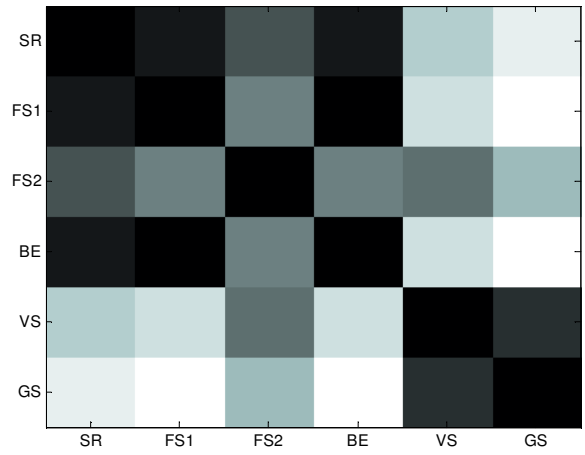


Figure S10: Spearman correlation of the features ranking according to the various selection methods

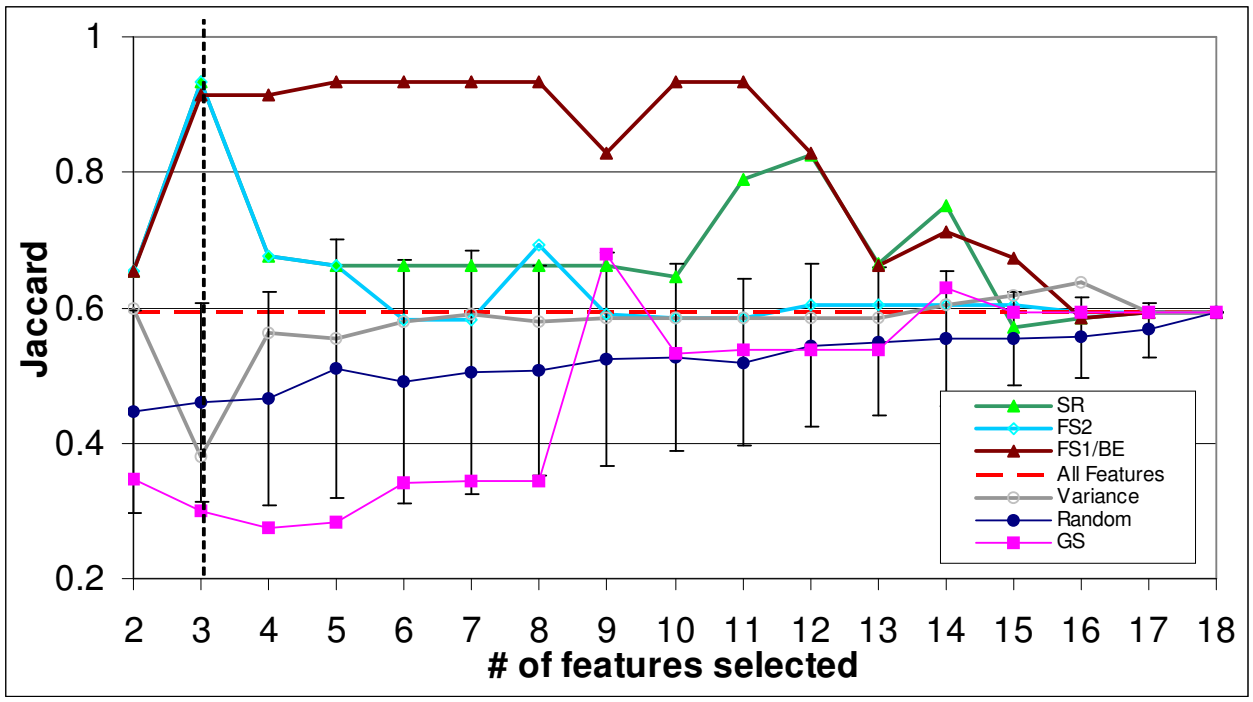


Figure S11: The quality of the various selection method of virus dataset (evaluation done by QC algorithm)

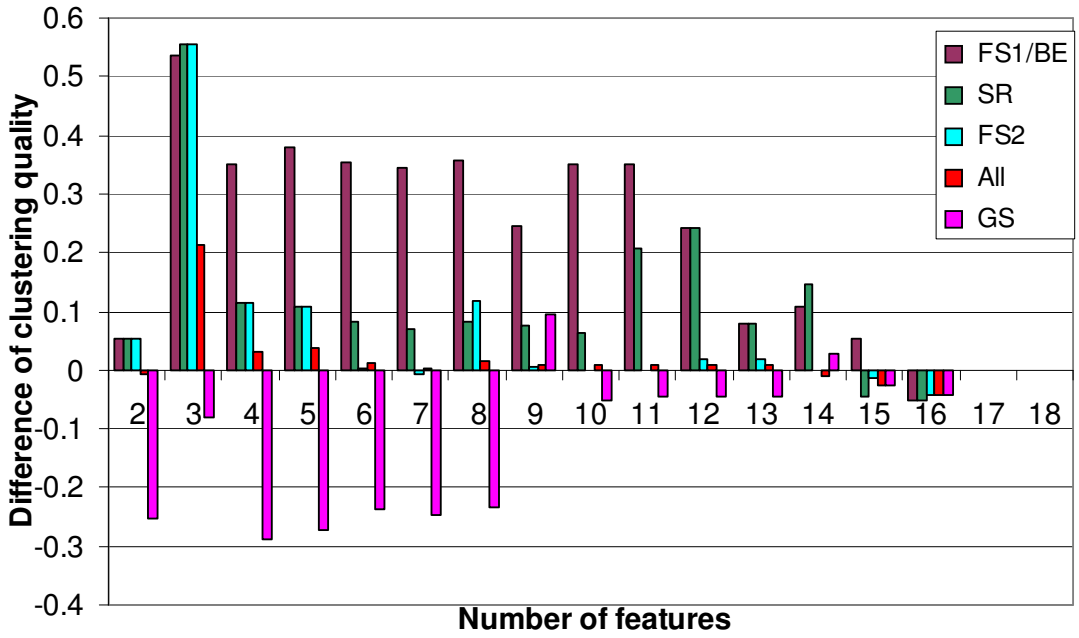


Figure S12: Difference in clustering quality of the Virus dataset, by selection according to the various selection methods. Displayed is the difference from the variance selection (VS).

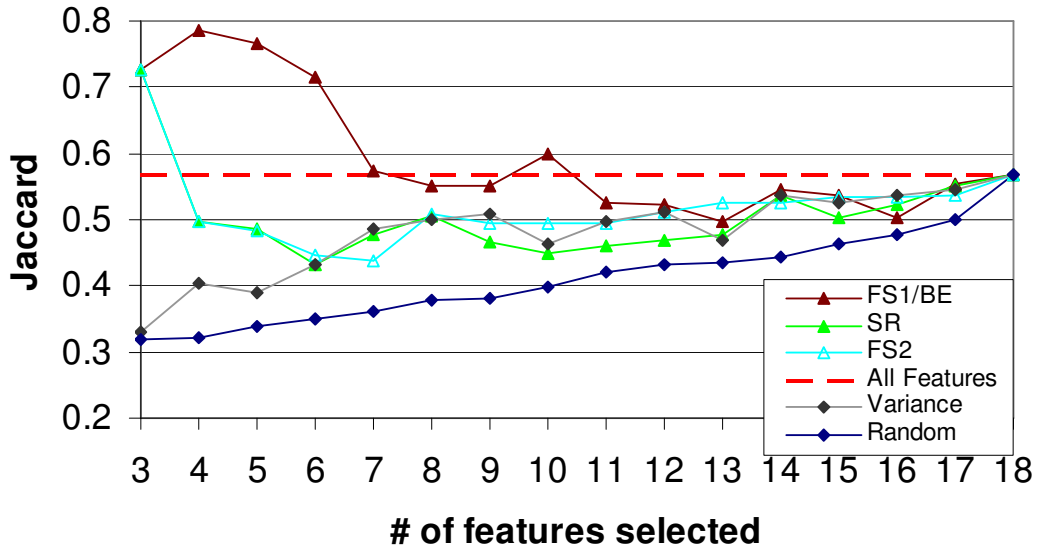


Figure S13: The quality of the various selection method of virus dataset (evaluation done by K-Means algorithm

3.7.2 The MLL Leukemia dataset of Armstrong et al., 2002

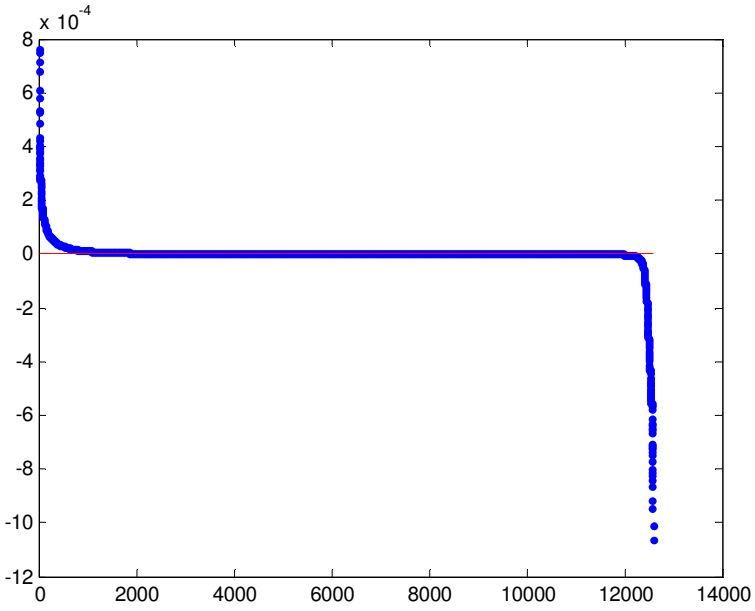


Figure S14: Simple Ranking of the MLL dataset

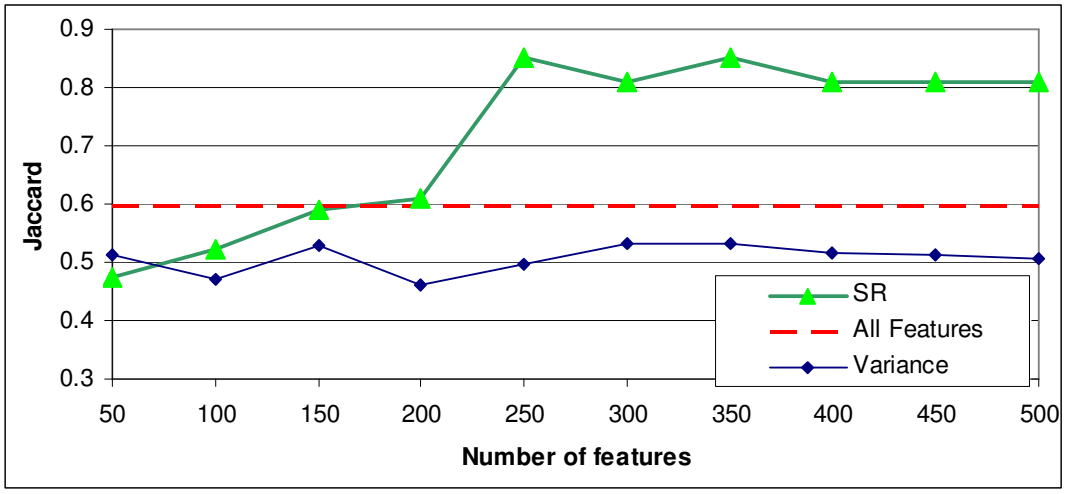


Figure S15: The quality of the various selection method of MLL dataset (evaluation done by QC algorithm

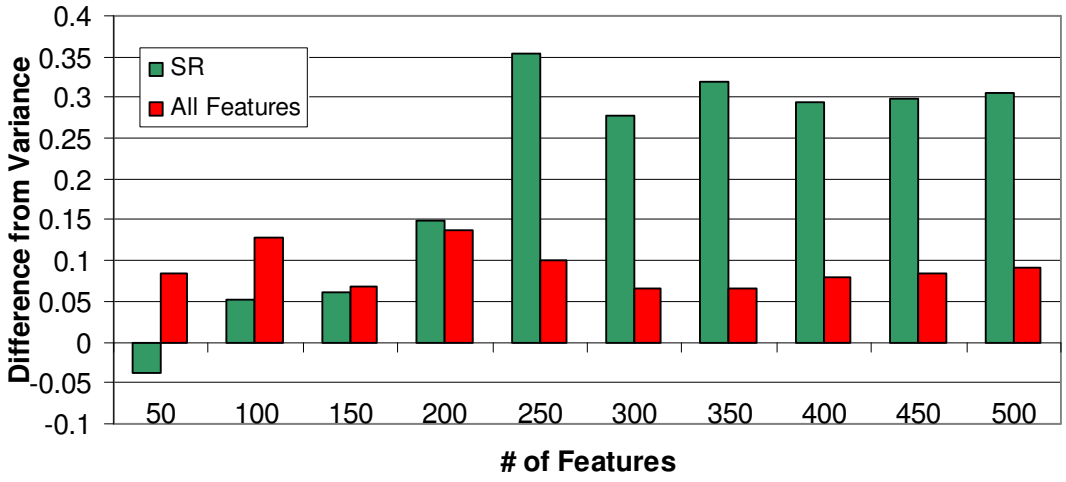


Figure S16: Difference in clustering quality of the MLL dataset, by selection according to the various selection methods. Displayed is the difference from the variance selection (VS).

3.7.3 The Leukemia dataset of Golub et al. 1999

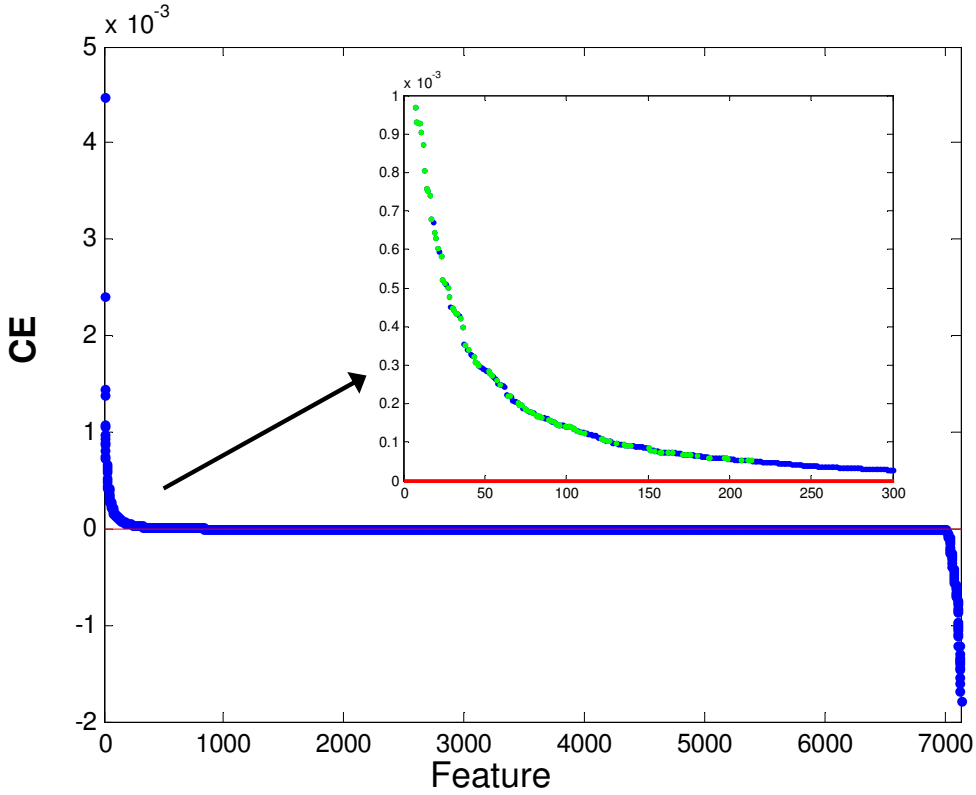


Figure S17: CE of the 7129 genes of the virus dataset. The inset provides zoom into the highest-ranked 300 genes, with bright dots signifying the top 100 features according to the FS1 method.

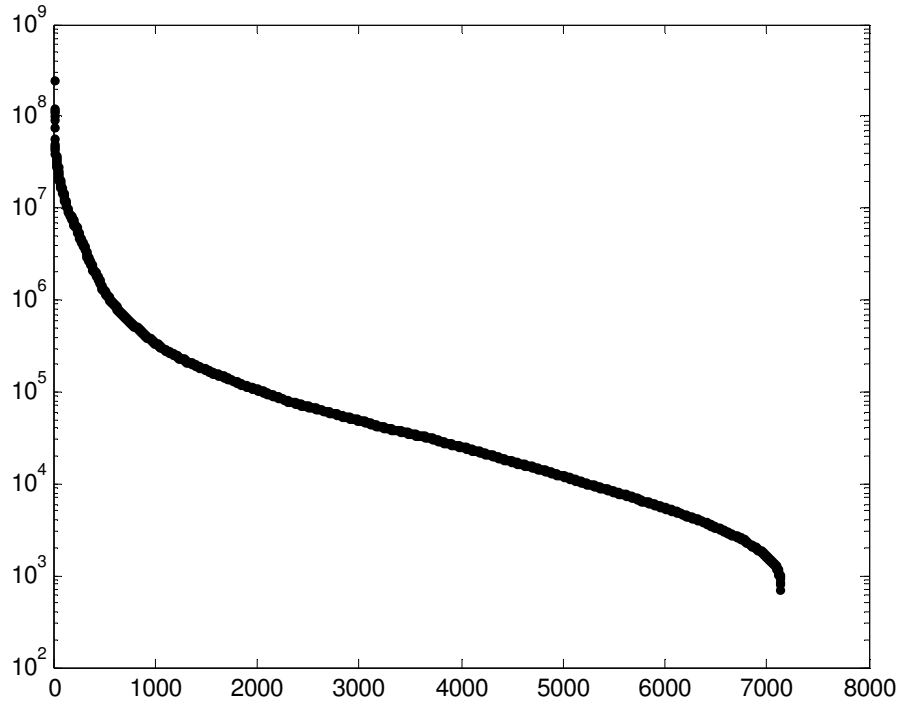


Figure S18: Log of the variance of the 7129 gene in the Leukemia dataset

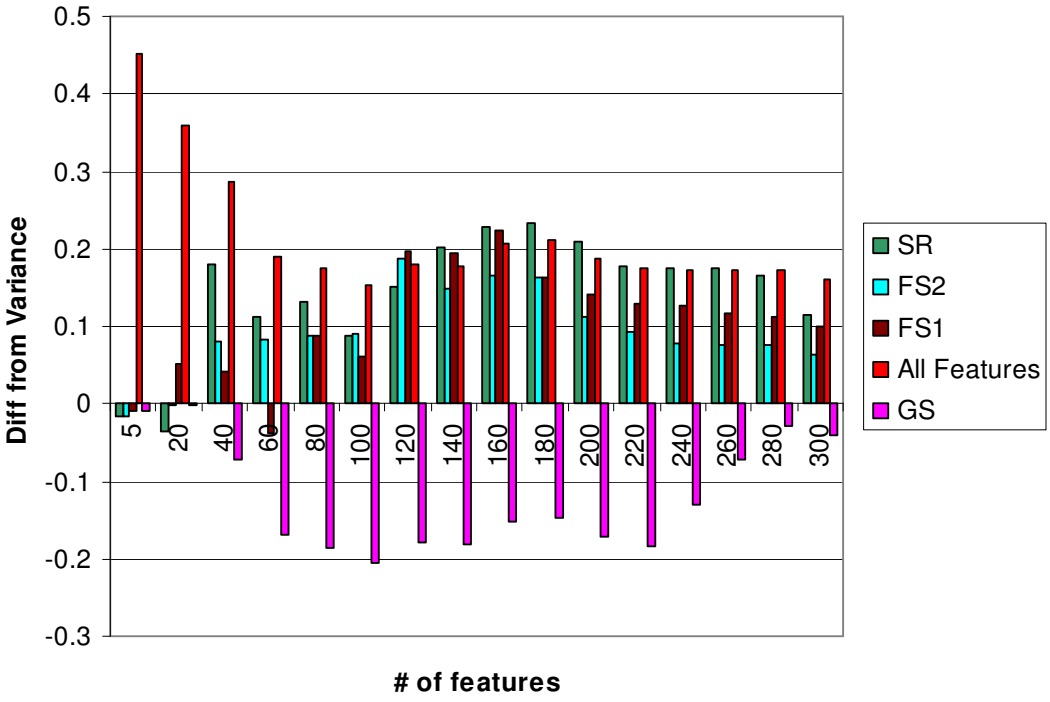


Figure S19: Difference in clustering quality of the Leukemia dataset, by selection according to the various selection methods. Displayed is the difference from the variance selection (VS).

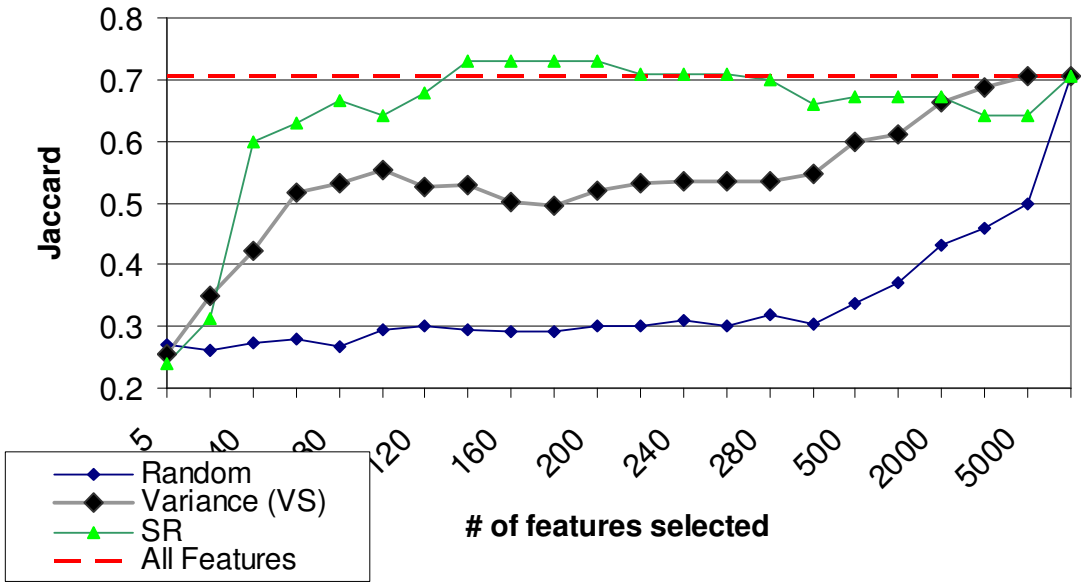


Figure S20: The quality of the various selection method of Golub dataset (evaluation done by QC algorithm

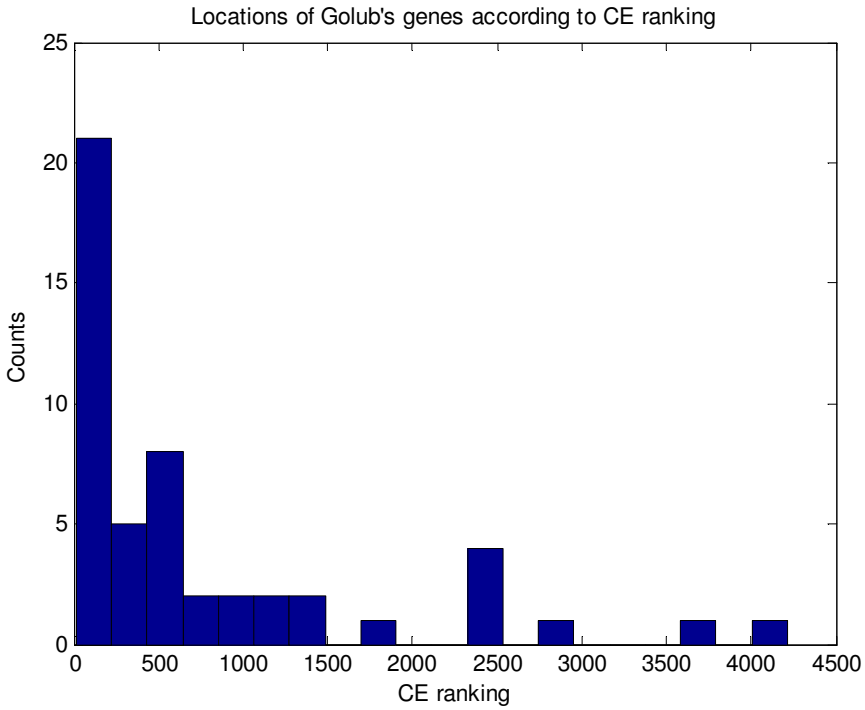


Figure S21: The location of Golub's 50 genes on the CE ranking graph

Chapter 4

UFFizi: A Generic Platform for Ranking Informative Features[7]

4.1 Introduction

The present information age is characterized by exponentially increasing data, e.g. in numbers of documents and in records of various kinds or biological data. Improved experimental techniques, such as high throughput methods in biology, allow for the measurement of thousands of features (genes) for each instance (single gene-expression microarray per patient). This leads to a flood of data, whose analysis calls for preprocessing in order to reduce noise and enhance the signal through dimensionality reduction. This is important for both enabling the application of various categorization techniques and allowing for biological inference from the data.

Dimensionality reduction algorithms are usually categorized as extraction or selection methods. Feature extraction transforms all features into a lower dimension space, while feature selection selects a subset of the original features. A benefit of the latter is the ability to attach meaning to the selected features. This is important both for exploration of the biological reality and for preparing a more concise experimental layout. The method to be studied here is categorized as feature selection.

It is customary to divide feature selection methods into two types: supervised, in which a target function is known and one tries to rank features or optimize some objective function relative to it, and unsupervised, in which one has no additional information regarding the instances. In practice, the abundance of unlabeled data or data that might posses multiple possible labeling, calls for an unsupervised approach.

While supervised feature selection methods are abundant [1], unsupervised methods are scarce, most of them tested on labeled data [2]. Nevertheless, unsupervised feature selection methods may play an important role even in supervised cases. Being unbiased by the labeling of the instances, unsupervised feature selection can be used as a preprocessing tool for supervised learning algorithms providing reduction of overfitting (for a comprehensive review we refer to [2]). As described in [3], feature selection from unsupervised data can be applied at three different stages: before, during and after clustering. Methods that operate before clustering are referred to as filter methods. Common methods of unsupervised feature filtering rank features according to either (1) their non-zero loadings in the first principal components [4] , (2) their normalized range,(3) entropy or (4) variance

[7] Based on the paper UFFizi: A Generic Platform for Ranking Informative Features, Assaf Gottlieb, Roy Varshavsky, Michal Linial and David Horn, Submitted.

of the feature as calculated from its values on all instances [2, 5]. All these methods estimate the importance of each feature independently of all others.

Our Unsupervised Feature Filtering (UFF) algorithm [6] differs from aforementioned methods in that it ranks features based on a criterion that involves all other features. It also provides a natural cutoff for selecting the number of features. We have also previously showed that UFF also selects stable feature sets under perturbations [7]. Our aim in this article is to introduce a new framework, based on the UFF. We (1) explore the properties of UFF and the features it selects, (2) introduce a faster approximate version, (3) suggest indicators for the ability to apply the method to certain datasets and (4) extend it by proposing a method called Unsupervised Instance Selection (UIS) for inspecting and eliminating potential outlier instances. A faster version of UFF, together with identification of indicators for the ability to apply the method to different datasets enables the implementation of UFF as a web-tool. The performance of the UFF is shown to surpass commonly used unsupervised filtering methods (e.g. variance, feature entropy) for the datasets used in this study. These findings are consistent with the findings reported in ([6].

In the Results section, we explore the properties of UFF on example datasets, introduce a faster algorithm for UFF and analyze which datasets can be evaluated successfully by the UFF method. We then describe the UDO method and provide biological insights on gene and microRNA expression from a wide range of diseased states.

4.1.1 List of abbreviations

UFF, Unsupervised Feature Filtering; SVD, Singular Value Decomposition; UIS, Unsupervised Instance Selection; CTD, Comparative Toxicogenomics Database

4.2 Methods

4.2.1 Datasets

We use three gene-expression microarray datasets with known labeling in order to demonstrate the performance of UFF. They were compiled from the online public repository of the National Center for Biotechnology Information/GenBank Gene Expression Omnibus (GEO) database [8], [9] . Data collections are: (i) Gene expression measurements taken from skin tissues including 7 normal skin tissues, 18 benign melanocytic lesions and 45 malignant melanoma [10] (series entry GSE3189); (ii) HIV dataset (series entry GSE6740), containing gene expression measurements from 20 CD4+ and 20 CD8+ T cells from HIV patients at different clinical stages; (iii) Hepatitis C (series entry GSE11190) containing gene expression measurements from 78 samples, comprising of 38 blood samples and 40 liver biopsy, before and after interferon treatment of Hepatitis C (19 blood samples

before and after the treatment, 21 and 19 liver biopsies before and after respectively). All these datasets are Affymetrix Human Genome U133A Array (Hepatitis C is a U133 plus 2.0 array).
In addition, we present results obtained from using UFF on The Cancer Genome Atlas (TCGA) gene-expression and microRNA (miRNA) expression datasets[11]. These datasets are comprised of samples taken from (i) glioblastoma multiforme (GBM) and (ii) ovarian serous cystadenocarcinoma (OV) patients. Gene-expression datasets are measured using Affymetrix Human Genome U133A Arrays and Agilent G4502A_07 platforms. miRNA expression is measured using Agilent Human miRNA Microarray Rel12.0 and Agilent 8 x 15K Human miRNA-specific platforms. Details of these datasets are specified in Table S1 in the supplementary material.

4.2.2 Unsupervised Feature Filtering (UFF)

UFF is based on an entropy measure applied to Singular Value Decomposition (SVD). Let A denote a matrix, whose elements Aij denote the measurement of feature *i* on instance *j*, e.g. expression of gene *i* under condition *j*. SVD decomposes the original matrix A into $A=USV^T$, where U and V are unitary matrices whose columns form orthonormal bases. The diagonal matrix S is composed of singular values (s_k) ordered from highest to lowest. SVD is a common technique in feature extraction. UFF uses the information contained in the singular values in order to select the features. Let q be the rank of the matrix (q=min(n,m), where n is the number of instances and m is the number of features). Using the singular values, s_k, one may define the normalized relative squared values ρ_k [12] [13]:

$$\rho_k = s_k^2 \Big/ \sum_{i=1}^{q} s_i^2 \qquad (5)$$

A dataset that is characterized by only a few high normalized singular values, whereas the rest are significantly smaller, reflects large redundancy in the data. On the other hand, non-redundant datasets lead to uniformity in the singular values spectrum. UFF exploits this property of the spectrum in order to measure how each feature *i* influences this redundancy, while favoring features which decrease redundancy. The score of a feature *i* is defined using a leave-one-out principle. A function *f* is calculated on the set of all singular values for the original matrix and for the corresponding set of the matrix without feature *i*. The difference in the values of *f* defines the score of each feature i. In this work, we use the SVD-entropy (*H*) as the function *f* [13] [14] (note that this 'Shannon'-like function does not use probabilities). The score of a feature can be thus regarded as its contribution to the SVD-entropy.

$$f \equiv H = -\frac{1}{\log(q)} \sum_{k=1}^{q} \rho_k \log(\rho_k) \qquad (6)$$

Other functions may be used instead of *H*. They have to be monotonic and vary from a maximum, when all singular values are equal, to a minimum when there is only one singular value bigger than zero. Two such functions that we tested are the negative value of sum of squares and the geometric mean. The results using these functions are very similar to those obtained using the SVD-entropy, hence we will not elaborate further on them.

Figure 1 displays the typical results after applying the UFF algorithm to the melanoma dataset (see the datasets subsection for description), and sorting the features according to the decreasing score of the UFF. Clearly, one can divide the features into three groups:

1. Features with positive score. These features increase the entropy.
2. Neutral features. These features have negligible influence on the entropy.
3. Negative score features. These features decrease the entropy.

We follow the Simple Ranking (SR) method of UFF, denoting positive score features (group 1) as features whose scores lie above the mean score + one std (upper dotted line in figure 2), negative score features (group 3) as features whose scores lie below the mean score - one std (lower dotted line) and neutral features (group 2) the rest. Note that most features fall into group 2, while groups 1 and 3 represent minorities. UFF [6] selects group 1 as containing the most relevant features. The rationale behind this selection is that, because these features increase the entropy, they decrease redundancy. Hence one may expect that instances will be better separated in the space spanned by these features. Further analysis of this group and its comparison with the two other groups is presented in the "properties of selected features" section.

.

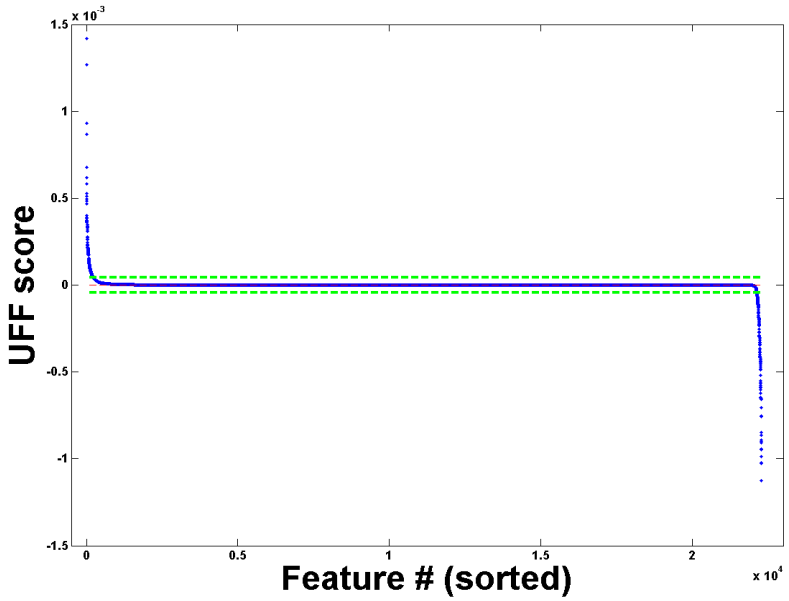


Figure. 1. UFF Scores of the 22283 genes of the melanoma dataset, ordered by decreasing scores. Dashed lines represent mean(score)±std(score).

In this paper, we follow the Simple Ranking (SR) method of UFF, selecting all positive score features (group 1). Alternative UFF methods suggested in [6] are not shown.

4.2.3 GO and Pathway Enrichment

Enrichment of Gene Ontology (GO), KEGG pathways and PubMed papers presented here were calculated using the DAVID [15], [16] and ToppGene tools [17]. Verifications were also done using other tools such as Ontologizer [18] and GO Tree Machine [19]

4.2.4 UFF Performance Validation

Clustering comparison between different unsupervised feature selection methods was performed using the widely used *k*-means clustering algorithm. In order to provide an unbiased comparison, all feature selection methods were tested with the same input parameter *k* (*k*=3 for the melanoma dataset, *k*=2 for the HIV dataset and *k*=4 for the Hepatitis-C dataset) for the *k*-means clustering algorithm with no additional preprocessing. The clustering was repeated 100 times for each feature selection method and each number of selected features.

Random selection was used to generate 100 different sets. Feature entropy was performed on each feature individually, using the same formalism as in equation 3. We used the Jaccard score [20] to measure the quality of the clustering relative to known labels.

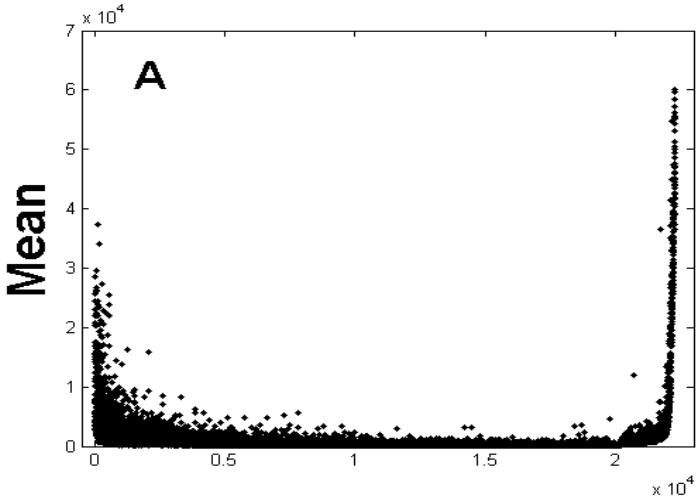
4.3 Results

4.3.1 Analyzing and Improving UFF

In this section, we present analysis of UFF selected features and provide improvements and extensions to the algorithm. The improvements include (i) Faster version of the algorithm and (ii) Addition of a criterion for assessing the quality of the results provided by UFF. We further extend the algorithm by introducing the Unsupervised Detection of Outliers (UDO).

4.3.1.1 Properties of Selected Features

We investigated the general properties of features selected by UFF, by studying their statistical properties. We demonstrate these properties on the melanoma gene expression dataset (see Methods). Figure 2 displays the mean (A) and variance (B) of all features (as measured on all instances), for the melanoma dataset. The features are ordered by their UFF rank, which is displayed in Figure 1. Dotted lines, denoting the mean (score) ± one standard deviation, supply the separation between the positive (group 1), neutral (group 2) and negative (group 3) score features (Methods). Most features belonging to the second (neutral) group possess low mean and variance. It is evident that both the positive score features and the negative score features have high mean (in general high absolute values of mean) and variance. This explains a major difference between UFF and the Variance Selection method: while UFF selects features from group 1, Variance Selection chooses features from both groups 1 and 3. It should be noted that if datasets of this nature (e.g. gene-expression) undergo standardizing operations, UFF selection may be meaningless.



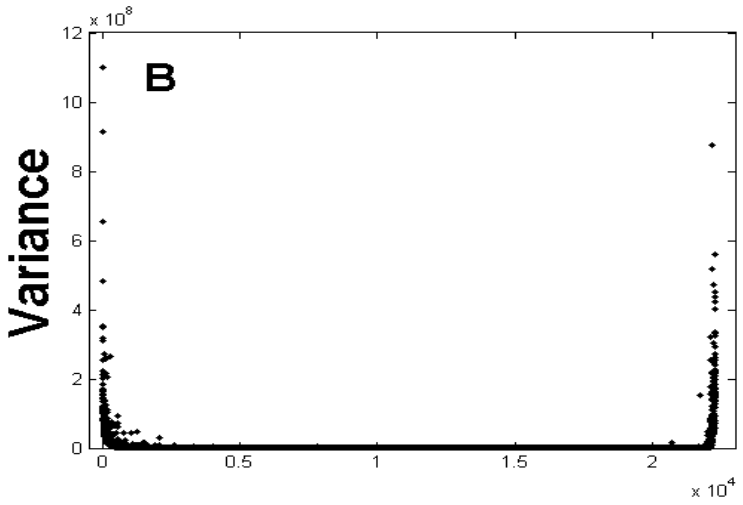


Figure. 2. (A) mean and (B) variance of the melanoma dataset (X axis refers to genes ordered according to UFF score).

An important difference between the positive (group 1) and negative (group 3) features is displayed in Figure 3. This figure shows the projection of typical positive and negative features (A and B, respectively) on the SVD eigenvectors (or principal components, PCs) of the original data matrix. Positive score features have more evenly distributed projections on the PCs relative to the negative score features, which project most strongly on the first PC. It is the latter property that explains the negative score: by preferring the leading principal component these features decrease SVD-entropy. We present in the Appendix a proof showing that when a feature lies only on the first PC, it is bound to have a negative score.

The differences in projection on the principal components between the positive and negative scored features, may provide an explanation for the difference between our approach and the sparse-PCA approach [4]. The latter selects genes that correlate mainly with the leading PC, while UFF prefers a wider distribution.

Finally we observe that negative score features have skewness close to zero and kurtosis close to three. Hence we conclude that negative score features possess wide Gaussian distributions, which can be regarded as baring no indicative signal over the instances. These noisy features are discarded by UFF but selected by Variance Selection, which explains their inferior results demonstrated in [6]

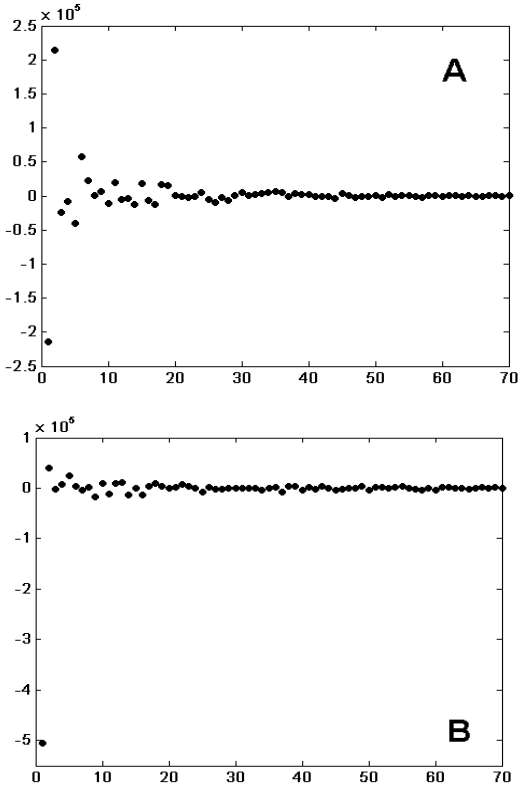


Figure 3. Projection on the 70 principal components of a typical - (A) positive score and (B) negative score - feature from the melanoma dataset. Note the outstanding value of PC1 in B.

4.3.1.2 Fast UFF

In order to obtain the UFF ranking of features one performs M times the SVD evaluation, where M is the number of features. This has the complexity of $O(M*min(N,M)^3)$ (see [6]). The data matrix A of M features by N instances is often represented by its SVD transformation $A=USV^T$, where U and V are unitary and S is the diagonal matrix of the singular values. The associated Gram matrix $C=A^TA$, of size NxN, can then be written as $C=VS^2V^T$, with eigenvalues that are the squares of the singular values of A and thus can be used directly to calculate the SVD-entropy. Removing a row from A, i.e. removing the feature f^k of length N, the Gram matrix C changes to

$$C \rightarrow C - \left(f^k \right)^T f^k \equiv C' \qquad (7)$$

We assume that removal of one feature can be regarded as a small perturbation, an assumption which generally holds for a large enough number of features. The singular values can be approximated by using the eigenvectors of the Gram matrix C on the new matrix C'. Plugging into equation (1), the changed SVD entropy is:

$$H\left(V'C'V'^{T}\right) \approx H\left(VC'V^{T}\right) = H\left(S^{2} - (Vf^{K})^{2}\right) \qquad (8)$$

An extended formulation is given in the Appendix.

This approximation reduces the complexity to $O(M*N^2)$ leading to considerable faster calculations. Table 1 compares the running times of fast UFF vs. regular UFF for three of the datasets used in this paper. As can be seen, the reduction in running time is substantial, allowing for an online computation.

The quality of the approximation lies in the assumption of small perturbations. In order to test whether this assumption holds for a given dataset, we inspect the SVD entropy of the matrix, defined to lie between 0 and 1 (see Methods). For most data-sets that we studied it is smaller than 0.1. Such a small value of the entropy guarantees that only a few eigenvalues (principal components) are of importance, and the removal of a single feature is indeed a small perturbation assuring the validity of the approximation (equation 2). In two of the studied datasets (GBM and OV microRNA) the SVD entropy is large (0.59 and 0.34 correspondingly), putting the approximation (equation 2) in doubt. In both cases one should therefore resort to the regular UFF calculation to obtain reliable results

Fast UFF allows for the analysis of much larger datasets. Moreover it enables incorporating this algorithm in a web-based tool. Computationally, it allows for a distributed evaluation of UFF scores, once the eigenvectors of the Gram matrix C are obtained. The calculation of the SVD entropy of the matrix is incorporated into the UFFizi web tool, initiating a warning when the results of the fast UFF might deviate substantially from the regular UFF.

4.3.1.3 When is UFF Applicable

While UFF works very well on many datasets, including most gene-expression data we have analyzed, we have found datasets where selection according to UFF is not effective. Figure 4 presents such an example using a dataset of pre-selected cell-cycle regulated genes. On such a dataset, UFF did not lead to improved clustering. We note that the distribution of score values in Figure 4 is somewhat different from Figure 1. In particular, group 2 features display large variance among their scores.

Working with more than twenty datasets from different domains, we have found measures that allow for separation between datasets on which UFF is effective from datasets in which it is not. One such measure is the normalized entropy of the squares of UFF scores. This, as well as another measure, is presented in the supplementary appendix. They allow for a prior estimate on whether UFF selected features should be used. These measures, formulated in the supplementary appendix, are incorporated into our web-tool, providing a confidence level for relying on UFF results.

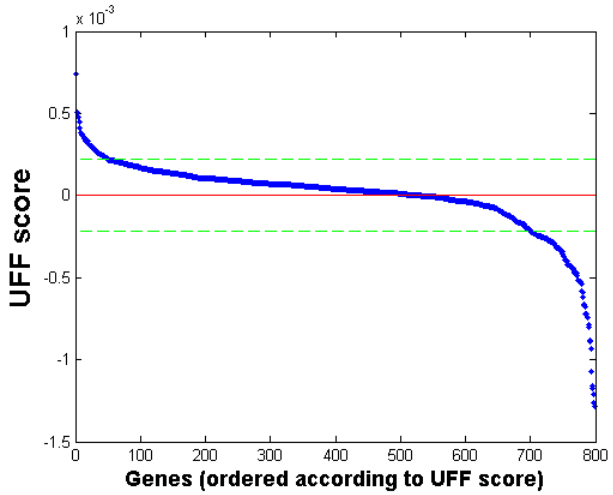


Figure. 4. UFF Scores of the Spellman cell-cycle dataset, ordered by decreasing UFF score.

4.3.1.4 Unsupervised Detection of Outliers (UDO)

Outliers are typically defined as instances that differ significantly from other instances in the data (for extensive surveys, see [21, 22]). Detecting such outlier instances may be desirable in certain cases, e.g. when there is a suspicion of faulty or unreliable measurements or for detecting rare events. A multitude of methods for unsupervised outlier detection have been proposed. Most relate to one of two approaches: (1) model based, in which a model is fit to the data and outliers are the ones deviating from the model [23, 24], (2) Distance-based methods, which find instances lying far from all instances, nearest instances, or nearby clusters [25-31]. We present here an alternative definition and a method to detect such outliers, based on the UFF framework.

The data-matrix A contains information on instances in terms of features and features in terms of instances, and the singular values are common to both. One may therefore consider a 'leave-one-out' measure applied to instances. This is the Unsupervised Detection of Outliers (UDO) method, to be studied here. UDO identifies instances that, when removed, decrease the entropy of the dataset and thus provide a more homogeneous dataset. Recognizing these entropy-increasing instances as outliers provides a natural definition for an “outlier-degree”. UDO attaches to each instance the amount of decrease of the SVD entropy, which is considered the global measure of the “outlier-degree” of each instance in the dataset. As in the UFF method, a threshold of one standard deviation (std) above the mean may be applied to assess the number of such outliers. UDO is a data-driven method, making no prior assumption regarding the distribution of the data such as model-based methods. It is not restricted by small sample size datasets which prohibit creation of valid distribution assessments. It is also different from distance-based outlier detection schemes in that it assesses the influence of instance removal on the entire dataset rather than the mere location in

feature space of the instance relative to other instances. In contrast to the Donoho-Stanhel estimator, which assesses the "outlier-degree" of an instance relative to one selected direction in feature space, UDO estimates it on all eigenvectors at once. UDO in this sense emphasizes directions along which other instances are relatively comparable. We note that in datasets of relatively low SVD entropy, the correlation between the UDO ranking and the popular outlier detection method of the k^{th}-NN ranking [29] is relatively high (0.61 and 0.82 for the melanoma and HIV datasets respectively, k=5). This can be explained by noting that removal of an instance in such datasets does not alter the leading eigenvectors substantially and UDO thus selects the high-entropy instances that reside mainly farthest along these eigenvectors. In high SVD entropy datasets (e.g. the two microRNA datasets in this paper), the correlation between the two different methods is essentially zero.
Since outlier defining criterion and the methods implementing them are intertwined, evaluation of each method turns often into subjective inspection of the outliers. We note that in the HIV dataset for which we have some clinical information, the first 4 selected instances (out of 5 selected by UDO) are samples of two individuals (containing both CD4+ and CD8+ T cells). The two leading outlier instances belong to the same individual, possessing an HIV infection at a very preliminary stage (~1 month), possibly explaining high divergence of measurements from individuals with longer periods of HIV infection.

4.3.2 Selected Datasets

In this section we present novel results obtained by applying UFF to gene-expression and microRNA (miRNA) expression datasets.

4.3.2.1 Melanoma – UFF selected genes

The melanoma dataset is used for demonstrating the different traits of UFF. Running UFF on this dataset, we obtain 231 genes. The top ranked genes include Stratifin, Keratin 14, Keratin 1 and Loricrin, mutation in which are related to skin cancer and other skin diseases [32-35]. Enrichment analysis includes terms having Bonferroni score<0.05. GO Enrichment analysis of the selected genes includes functions of biological processes such as ectoderm and epidermis development, homophilic cell adhesion, keratinocyte differentiation and melanin biosynthetic process. Cellular compartments enrichment includes intermediate filament, extracellular region and melanosome. Interestingly, GO molecular function enrichment show various metal ion binding, including copper, cadmium and calcium, all having relations to the tumor suppressor protein p53 [36-38]. Enriched pathways include cell communication, antigen processing and presentation and also breast cancer estrogen signaling. Human phenotype analysis reveals enrichment for palmoplantar hyperkeratosis, keratinization, skin

and integument abnormalities. The list of UFF selected genes is provided in supplementary Table S2. The full list of GO enrichment terms is provided in supplementary Table S3.

Talantov, et al. (2005) performed clustering analysis on this dataset, using a filtered list of 15,795 genes. They did not obtain perfect separation between melanoma and benign tumors or normal tissues (obtaining Jaccard score [20] of 0.74). Using UFF selected genes and the Quantum Clustering algorithm [39], we were able to correctly split melanoma from benign tissues, while identifying two clusters in the melanoma samples similar to the ones identified by [10] (Jaccard score of 0.85)32 of UFF selected genes appear also in the 439 differentially expressed genes of [10] (p-value = e^{-12}) and 10 out of 33 differentially expressed genes with high fold change (p-value<e^{-12}).

Figure 5 compares the clustering results in terms of Jaccard score using UFF selected genes for different thresholds, with genes selected using variance, feature entropy and random selection and using all the genes (see Methods). It is evident that UFF features provide better clustering results than either selection method or compared to using all the genes for all thresholds (with an exception for the top 10 genes, where variance selection has slightly better Jaccard score). Error bars were removed for clarity. Supplementary figure S1 displays the same comparison with error bars. Quantum Clustering results are provided in supplementary Table S4.

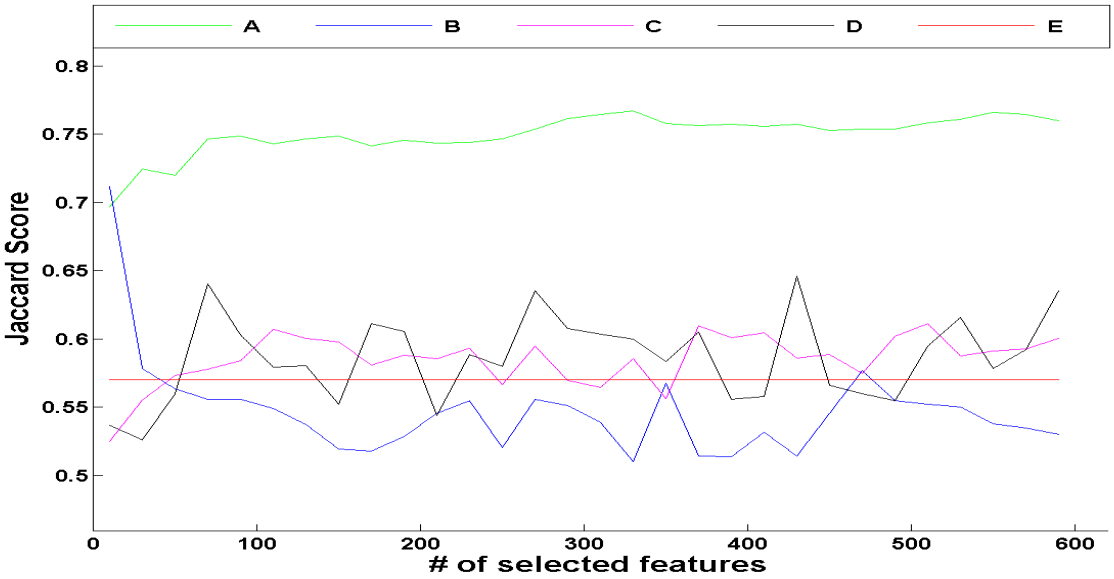


Figure. 5. Mean Jaccard scores of clustering results for different selection methods on the melanoma dataset. Tested methods include (A) UFF, (B) Variance, (C) Feature entropy, (D) Random selection and (E) All features.

4.3.2.2 HIV – UFF selected genes

Next we explored the HIV dataset. UFF selected 179 genes, enabling us to cluster the CD4+ and CD8+ samples into separate clusters with only one misclassification. In comparison, when we clustered the samples using all the genes 2 misclassifications were obtained. In the top ranking genes we find mostly hemoglobin units, but also the specific CD4+ HIV related protein defensin [40] and the CD8+ HIV related CD8 antigen [41]. GO enriched biological processes for the 179 selected genes (Bonferroni<0.05) include immune system process, immune response, cellular defense response, antigen processing and presentation of peptide antigen via MHC class I and class II. Cellular compartments are enriched for the MHC class I and II protein complexes. Non trivial enriched pathways include Graft-versus-host disease, natural killer cell mediated cytotoxicity and type I diabetes (Bonferroni<10^{-6}). The selected genes involved in the type I diabetes pathway are usually in direct connection with either CD4+ or CD8+ T-cells. This connection is strongly support by literature text mining (not shown). The list of selected genes is provided in supplementary Table S2. Enriched terms are provided in supplementary Table S3.

Similar to figure 5, supplementary figure S2 displays the performance of clustering the HIV instances using different gene sets, selected by various unsupervised feature selection methods, random selection and using all the genes. The performance of UFF surpasses all other methods in terms of clustering results (see Methods).

4.3.2.3 Chronic Hepatitis -C – UFF selected genes

The CHC database is intended for inspecting results of chronic hepatitis C (CHC) treatment with interferon. UFF selected 513 genes. Using these selected genes, we were able to separate perfectly pre-interferon and post interferon blood samples. Liver biopsies, however, were clustered according to sample origin instead of pre and post interferon treatment. The clustering results are different when using all the genes; in this case, liver samples could not be separated at all and blood samples typically split into different clusters. This is displayed in Figure 6. The relevance of the gene selected is demonstrated by the GO enrichment scheme. The GO cellular compartment contains various lipoprotein particles (high-density, plasma, spherical high-density, triglyceride-rich, very-low-density and intermediate-density). Biological process enrichment includes lipid metabolic process, along with regular defense system terms, such as acute inflammatory response, response to wounding and response to xenobiotic stimulus and metabolism of xenobiotics by cytochrome P450 pathway, possibly related to the Interferon treatment [42]. An enriched human phenotype is generalized amyloid deposition, which is reported to relate to hepatitis C [43]. Finally, using the Comparative Toxicogenomics Database (CTD) the UFF selected genes are enriched for Hepatitis and the related immune complex diseases. UFF selected genes and enrichment analysis are provided

in supplementary tables S2 and S3 respectively. Clustering results appear in supplementary Table S4.

Supplementary figure S3 compares the performance of clustering the Hepatitis-C instances using UFF selected genes with gene sets selected by various unsupervised feature selection methods, random selection and using all the features. The performance of UFF again tops other methods in terms of clustering results.

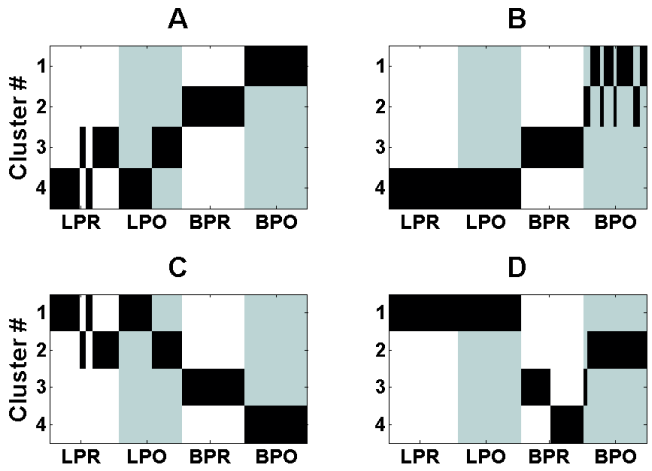


Figure 6. Clustering of the 78 samples of Hepatitis C dataset, relative to known labeling. Y-axis denotes cluster number and X-axis denotes division into pre-interferon liver biopsy (LPR), post-interferon liver biopsy (LPO), pre-interferon blood sample (BPR) and post-interferon blood sample (BPO). Clustering was performed using both *k*-means (k=4) using UFF selected genes (A) and using all genes (B) and by using Quantum Clustering using UFF selected genes (C) and using all genes (D)

4.3.2.4 Glioblastoma – UFF selected genes

We present results on glioblastoma multiforme (GBM) from The Cancer Genome Atlas (TCGA) project. We selected features from each platform independently, due to the difference between experiments, allowing for identification of genes that differentiate between different platforms, rather than different instance type (UFF was applied to AgilentG4502A_07_1 and AgilentG4502A_07_2 separately, to avoid selection of genes that allows perfect separation of the two platforms). The unsupervised approach displays its full strength in this case, since we do not have access to additional sample information on these datasets.

Based on UFF selected genes, we clearly identify clustering of the instances in each dataset into a small number of groups. As clinical details of the subjects are not specified, we cannot link these

clusters to known labels. An example of the clustering results for one of the GBM datasets is displayed in Figure 7. Clustering results of selected datasets are found in supplementary Table S4.

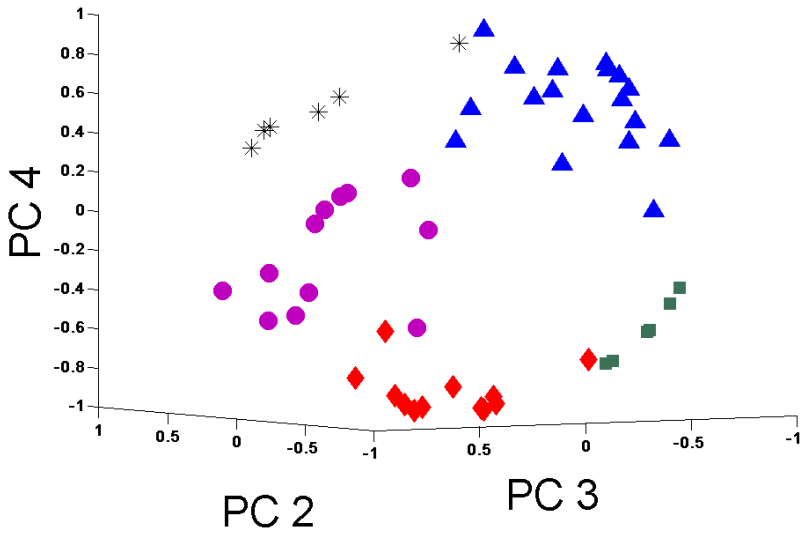


Figure 7. Clustering of 54 samples of GBM Agilent G4502A_07_1.4.2.0 array, colors and shapes denote different clusters. Image displays projection on principal components 2-4

There are variations between the number of genes selected on Agilent and Affymetrix gene expression platforms (563 and 731 genes for Agilent 1 and 2 platforms, while only 140 for Affymetrix).

We focus on the list of 44 genes, which are common to both platforms. 13 genes from this list also appear in the list of top 100 primary glioblastoma-associated genes expressed at higher levels compared with normal brain tissue [44]. We note also that 3 out of 4 patented markers for glioblastoma (patent #7115265) appear on this common list (the 4th marker, ABCC3, appears in genes selected from Agilent 2 platform). The top 10 genes from this list, in terms of minimal UFF rank, are displayed in Table 3. Supplementary Table S5 provides detailed explanations on relations to cancer biomarkers. UFF selected genes and the 44 common genes appear in supplementary Table S2.

Although Agilent and Affymetrix datasets show high variance in the number of genes selected by UFF, the highest GO enrichment terms are common to both. Both show high GO enrichment of general biological processes such as regulation of multicellular organismal process, cell proliferation and nervous system development (Bonferroni<0.05) and nervous system development in Affymetrix, (FDR<0.05, but Bonferroni <0.1). UFF selected genes on Affymetrix also show inflammatory response while UFF selected genes of Agilent are enriched for cell adhesion. Both platforms are also

enriched for cellular compartment of extracellular matrix and both were highly enriched for 'signal peptide' and 'secreted' (Bonferroni<0.0005) based on UniProt keywords. UFF selected genes on both platforms are enriched for molecular function of protein and receptor binding, which includes various ligands such as polysaccharide, heparin and neuropeptide hormone activity binding (Agilent platform), and lipid and ferric iron binding (Affymetrix platform). Enrichment analysis is provided in supplementary Table S3.

Table 2. Top 10 ranked genes, selected on all platforms of glioblastoma multiforme. Genes with asterisk appear on the list of [44]. N.D = Not Determined.

Gene name	Minimal UFF rank across platforms	Related to Cancer Biomarkers
RPS4Y1	1	N.D
SEC61G	1	Yes
POSTN (*)	2	Yes
ECOP	7	Yes
TMSL8 (*)	9	N.D.
SERPINA3 (*)	10	Yes.
COL1A2 (*)	12	Yes
NPTX2	13	Yes
TIMP1 (*)	14	Yes
VSNL1	17	Yes

4.3.2.5 Ovarian Serous Cystadenocarcinoma – UFF selected genes

We performed similar analysis of the glioblastoma multiforme (GBM) datasets on the ovarian serous cystadenocarcinoma (OV) dataset from TCGA . UFF selects 669 and 998 genes from Agilent and Affymetrix platform datasets respectively. GO enrichment analysis reveals that UFF selected genes expose very similar GO terms as UFF selected genes on GBM.

The first interesting exception is cellular compartment enrichment in which OV shows enrichment for collagen and fibril, which are identified as predictors for ovarian cancer [45], [46]. An enrichment term which includes arthritis and osteoarthritis is of special interest, as the former was postulated as a marker for ovarian cancer [47], while the later has not been determined. Finally, enriched diseases show stomach and breast neoplasms. Enrichment analysis is provided in supplementary Table S3. Clustering of the samples according to the UFF selected genes is provided in supplementary Table S4.

190 genes are common to both Agilent and Affymetrix platforms. Table 3 lists the top 10 common genes in terms of minimal UFF rank. Supplementary Table S5 provides detailed explanations for Table 3. List of UFF OV selected genes and the 190 platform-shared genes are provided in supplementary Table S2.

Table 3. Top 10 ranked genes, selected on all platforms of ovarian serous cystadenocarcinoma. N.D = Not Determined.

Gene name	Minimal UFF rank across platforms	Related to Cancer Biomarkers
IGF2	1	Yes
HOXA4	2	Yes
POSTN	3	Yes
LMO3	5	Yes
ZIC1	7	Yes
HOXA9	8	Yes
PCP4	8	N.D
OVGP1	9	Yes
PON3	9	N.D
CXCL1	10	Yes

7 of the UFF selected genes are common to both GBM and OV. These are POSTN, NPTX2, GJA1, NNMT, CSRP2, SCG5 and HSPA1A, all of them related to cancer biomarkers. Supplementary table S2 provides further details on relation of these 7 common genes to cancer biomarkers. Note that POSTN appears in the top 10 selected genes in both GBM and OV datasets.

4.3.2.6 Selected miRNA for GBM and OV

We also report UFF selected microRNAs (miRNA) from TCGA microarrays for the glioblastoma (GBM) and ovarian (OV) cancers. There are 534 miRNAs in GBM, taken from 325 samples and 799 miRNAs taken from 295 OV samples. UFF selected 43 and 63 miRNAs in GBM and OV respectively.

Almost all of the UFF selected miRNAs are human miRNAs (hypergeometric p-value=0.003 and 0.05 for GBM and OV respectively). The selected miRNAs for GBM and OV are enriched in comparison to [48] list of up or down-regulated miRNAs relative to normal tissue (15 and 20 genes, corresponding to p-values of $7*10^{-5}$ and $9*10^{-6}$ for GBM and OV respectively). In comparison, negative entropy miRNAs are not enriched relative to this list.

12 of the selected miRNAs appear in both GBM and OV tumors. They are listed in Table 4. Supplementary Table S6 provides further details on relation of these miRNAs to cancer biomarkers. Selected miRNAs for GBM and OV are also listed in supplementary table S6.

Table 4. MicroRNAs selected by UFF, common to GBM and OV.
[1] up or down-regulated microRNAs relative to normal tissue according to {Lee, 2008 #53}
[2] MicroRNAs that affect the properties of cancer cells according to {Lee, 2008 #53}
[3] down-regulated in ovarian cancer {Lee, 2008 #53}
[4] Differentially expressed miRNAs in ovarian cancer tissues and cell lines {Dahiya, 2008 #230}.
N.D = Not Determined.

microRNA	Minimal UFF	Related to Cancer

	rank	Biomarkers
hsa-mir-181a [1]	3	Yes
hsa-mir-363	4	N.D
hsa-mir-210 [2]	6	Yes
hsa-mir-451	7	Yes
hsa-mir-10a	7	Yes
hsa-mir-31[1]	8	Yes
hsa-mir-196a [1]	8	Yes
hsa-mir-145* [2,3]	10	Yes
hsa-mir-135b [1]	11	Yes
hsa-mir-10b [1,2,4]	11	Yes
hsa-mir-10b* [1,2,4]	11	Yes
hsa-mir-31* [1]	12	Yes
hsa-mir-424 [4]	18	Yes
hsa-mir-155 [1,4]	20	Yes
hsa-mir-222 [1,2]	25	Yes
hsa-mir-30a* [1,4]	26	Yes
hsa-mir-517*	31	N.D

4.4 Conclusions

We present an improved method, and a new web tool, that enable users to benefit from the power of UFF, an unsupervised approach that scores and ranks each feature according to its influence on the singular values distribution.

A statistical characterization of the selected features shows that our method selects features of high variance (over instances), but only those that do not have large correlation only with the first principal component. It turns out that thus we ignore noisy features that have Gaussian distributions. The strength of our method lies in selecting features that both capture inherent clustering of the instances and possess high variance. The combination of the two is significant in the case of biological datasets such as expression microarrays.

By studying various empirical datasets and evaluating different scoring functions we show that our approach is generic, and can identify the subset of relevant features. In contradistinction to other methods we can estimate the size of the group of selected relevant features. Furthermore, we present a novel approximation method, enabling significantly faster calculation of the UFF feature scores.

UFF is a heuristic method which exposes its strength in realistic applications. Nevertheless, not all datasets are amenable to feature selection by UFF. We propose criteria for deciding when UFF

application is effective. This information is also provided in the online UFF tool. We further extend the capabilities of UFF by introducing the Unsupervised Detection of Outliers (UDO) method. UDO provides a novel definition of an "outlier-degree" of an instance and identifies such outliers in the dataset. This enables the researcher to detect rare events in the dataset or filter faulty instances before proceeding with further analysis.

Finally, we analyze various gene expression and microRNA expression datasets to show the strength of our approach and to expose interesting findings on these datasets with possible biological relevance.

Web tool: http://adios.tau.ac.il/UFFizi

4.5 Acknowledgements

We thank Alon Kaufman and Nati Linial for stimulating discussions and suggestions. RV is a fellow member of the Sudarsky Center for Computational Biology.

4.6 References

1. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507-2517.
2. Guyon I, Elisseeff A: **An Introduction to Variable and Feature Selection**. *Journal of Machine Learning Research* 2003, **3**:1157--1182.
3. Dy JG, Brodley CE: **Feature Selection for Unsupervised Learning**. *J Mach Learn Res* 2004, **5**:845-889.
4. Zou H, Hastie T, Tibshirani R: **Sparse Principal Component Analysis**. *Journal of Computational and Graphical Statistics* 2006, **15**(2):265-286.
5. Herrero J, Diaz-Uriarte R, Dopazo J: **Gene expression data preprocessing**. *Bioinformatics* 2003, **19**(5):655-656.
6. Varshavsky R, Gottlieb A, Linial M, Horn D: **Novel Unsupervised Feature Filtering of Biological Data**. *Bioinformatics* 2006, **22**(14):e507-513.
7. Varshavsky R, Gottlieb A, Horn D, Linial M: **Unsupervised feature selection under perturbations: meeting the challenges of biological data**. *Bioinformatics* 2007, **23**(24):3343-3349.
8. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**(1):207-210.
9. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles--database and tools update** *Nucleic Acids Res* 2007, **35**:D760-D765.
10. Talantov D, Mazumder A, X.Yu J, Briggs T, Jiang Y, Backus J, Atkins D, Wang Y: **Novel Genes Associated with Malignant Melanoma but not BenignMelanocytic Lesions**. *Clin Cancer Res* 2005, **11**(20).
11. **The Cancer Genome Atlas, http://tcga.cancer.gov/**.
12. Wall M, Rechtsteiner A, Rocha L: **Singular Value Decomposition and Principal Component Analysis**. In: *A Practical Approach to Microarray Data Analysis.* Edited by Berrar D, Dubitzky W, Granzow M: Kluwer; 2003: 91-109.
13. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *PNAS* 2000, **97**(18):10101-10106.
14. Devijver PA, Kittler J: **Pattern recognition : a statistical approach**. Englewood Cliffs, N.J: Prentice-Hall; 1982.
15. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources.** *Nature Protoc* 2009, **4**(1):44-57.

16. Dennis.Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempick RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biology* 2003, **4**(P3).
17. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization**. *Nucleic Acids Res* 2009, **37**:W305-W311.
18. Robinson PN, Wollstein A, Böhme U, Beattie B: **Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology**. *Bioinformatics* 2004, **20**(6):979–981.
19. Zhang B, Schmoyer D, Kirov S, Snoddy J: **GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies**. *BMC Bioinformatics* 2004, **5**(16).
20. Jaccard P: **Nouvelles recherches sur la distribution florale.** *Bul Soc Vaudoise Sci Nat* 1908, **44**:223–270.
21. Hodge V, Austin J: **A Survey of Outlier Detection Methodologies** *Artificial Intelligence Review* 2004, **22**(2):85-126.
22. Zhang Y, Meratnia N, Havinga P: **A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets**. *Technical Report TR-CTIT-07-79, Centre for Telematics and Information Technology, University of Twente, Enschede* 2007.
23. Guyon I, Matic N, Vapnik V: **Advances in knowledge discovery and data mining**: American Association for Artificial Intelligence Menlo Park, CA, USA 1996.
24. Yamanishi K, Takeuchi J-i, Williams G, Milne P: **On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms** *Data Mining and Knowledge Discovery* 2004, **8**(3):275-300.
25. Donoho DL, Gasko M: **Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness**. *Ann Statist* 1992, **20**(4):1803-1827.
26. Donoho DL: **Breakdown properties of multivariate location estimators**. PhD qualifying paper, Harvard University; 1982.
27. Stahel WA: **Breakdown of Covariance Estimators**. *Research Report 31, Fachgruppe für Statistik, ETH Zürich* 1981.
28. Maronna RA, Yohai VJ: **The Behavior of the Stahel-Donoho Robust Multivariate Estimator**. *Journal of the American Statistical Association* 1995, **90**(429):330-341.
29. Ramaswamy S, Rastogi R, Shim K: **Efficient algorithms for mining outliers from large data sets**. *Proceedings of the ACM SIGMOD Conference* 2000, **29**(2):427 - 438.
30. Breunig MM, Kriegel H-P, Ng RT, Sander J: **LOF: Identifying Density-Based Local Outliers**. *ACM SIGMOD conference* 2000, **29**(2):93-104.
31. Zoubi MdBA-: **An Effective Clustering-Based Approach for Outlier Detection**. *European Journal of Scientific Research* 2009, **28**(2):310-316.
32. Herron BJ, Liddell RA, Parker A, Grant S, Kinne J, Fisher JK, Siracusa LD: **A mutation in stratifin is responsible for the repeated epilation (Er) phenotype in mice**. *Nature Genetics* 2005, **37**:1210 - 1212.
33. Chan Y, Anton-Lamprecht I, Yu QC, Jäckel A, Zabel B, and JPE, Fuchs E: **A human keratin 14 "knockout": the absence of K14 leads to severe epidermolysis bullosa simplex and a function for an intermediate filament protein.** . *Genes & Dev* 1994, **8**:2574-2587.
34. Rothnagel JA, Dominey AM, Dempsey LD, Longley MA, Greenhalgh DA, Gagne TA, Huber M, Frenk E, Hohl D, Roop DR: **Mutations in the rod domains of keratins 1 and 10 in epidermolytic hyperkeratosis.** *Science* 1992, **257**:1128-1130.
35. Maestrini E, Monaco AP, McGrath JA, Ishida-Yamamoto A, Camisa C, Hovnanian A, Weeks DE, Lathrop M, Uitto J, Christiano AM: **A molecular defect in loricrin, the major component of the cornified cell envelope, underlies Vohwinkel's syndrome**. *Nature Genetics* 1996, **13**:70-77.
36. Verhaegh G, Richard M, Hainaut P: **Regulation of p53 by metal ions and by antioxidants: dithiocarbamate down-regulates p53 DNA-binding activity by increasing the intracellular level of copper**. *Mol Cell Biol* 1997, **17**(10):5699-5706.
37. MéplanDagger C, Mann K, Hainaut P: **Cadmium Induces Conformational Modifications of Wild-type p53 and Suppresses p53 Response to DNA Damage in Cultured Cells**. *J Biol Chem* 1999, **274**(44):31663-31670.
38. Metcalfe S, Weeds A, Okorokov AL, Milner J, Cockman M, Pope B: **Wild-type p53 protein shows calcium-dependent binding to F-actin**. *Oncogene* 1999, **18**(14):2351-2355.

39. Horn D, Gottlieb A: **Algorithm for data clustering in pattern recognition problems based on quantum mechanics**. *Physical Review Letters* 2001, **88**(1).
40. Theresa L. Chang JV, Jr., Armando DelPortillo and Mary E. Klotman: **Dual role of α-defensin-1 in anti–HIV-1 innate immunity**. *J Clin Invest* 2005, **115**(3):765-773.
41. Chu F, Tsang PH, Robez JP, J.I.Wallace, J.G.Bekesi: **Increased spontaneous release of CD8 antigen from CD8+ cells reflects the clinical progression of HIV-1 infected individuals**. *Int Conf AIDS* 1989, **5**(431).
42. Hodgson PD, Renton KW: **The role of nitric oxide generation in interferon-evoked cytochrome P450 down-regulation**. *The role of nitric oxide generation in interferon-evoked cytochrome P450 down-regulation* 1995, **17**(12):995-1000.
43. Barsoum RS: **Hepatitis C virus: from entry to renal injury—facts and potentials**. *Nephrology Dialysis Transplantation* 2007, **22**(7):1840-1848.
44. Tso C-L, Shintaku P, Chen J, Liu Q, Liu J, Chen Z, Yoshimoto K, Mischel PS, Cloughesy TF, Liau LM *et al*: **Primary Glioblastomas Express Mesenchymal Stem-Like Properties**. *Mol Cancer Res* 2006, **4**:607.
45. Santala M, Simojoki M, Risteli J, Risteli L, Kauppila A: **Type I and Type III Collagen Metabolites as Predictors of Clinical Outcome in Epithelial Ovarian Cancer**. *Clinical Cancer Res* 1999, **5**:4091-4096.
46. Santala M, Risteli J, Risteli L, Puistola U, Kacinski BM, Stanley ER, Kauppila A: **Synthesis and breakdown of fibrillar collagens: concomitant phenomena in ovarian cancer.** *Br J Cancer* 1998, **77**(11):1825-1831.
47. Martorell EA, Murray PM, Peterson JJ, Menke DM, Calamia KT: **Palmar fasciitis and arthritis syndrome associated with metastatic ovarian carcinoma: a report of four cases.** *J Hand Surg* 2004, **29**(4):654-660.
48. Lee YS, Dutta A: **MicroRNAs in cancer**. *Annual Review of Pathology: Mechanisms of Disease* 2008, **4**:199-227.
49. Hellman-Feynmann: **theorem of quantum mechanical forces was originally proven by P. Ehrenfest, Z. Phys. 45, 455 (1927), and later discussed by Hellman (1937) and independently rediscovered by Feynman (1939)**. 1927.
50. Hellman H: **Einfuhrung in die Quantenchemie**. Leipzig and Vienna: Deuticke; 1937.
51. Feynman R, P: **Forces in Molecules**. *Physical Review* 1939, **56**:340 - 343

4.7 Appendix

4.7.1 Connection between projection on first principal component and negative entropy score

One can prove that in the extreme case, where a feature is lying only on the first PC, it is bound to have a negative score. We shall now prove it for the SVD-entropy function. This proof can be extended to cover also the alternative measures mentioned in section 4.2.2.

Starting with the positive-definite Gram matrix *C*, defined as

$$C = A^T A = VS^2V^T \qquad (9)$$

for the data matrix A of M features by N instances (where, without loss of generality we assume N≤M). We use the eigenvalues of the Gram matrix, defined by $c_i = s_i^2$ to define:

$$\rho_i = \frac{c_i}{T},\ T = \sum_{j=1}^{N} c_j,\ K = -\sum_{j=1}^{N} c_j \log(c_j) \qquad (10)$$

T is positive definite. SVD entropy can be related to *K* through

$$H = -\sum_{i=1}^{N} \rho_i \log(\rho_i) = \frac{K}{T} + \log\left(T\right) \qquad (11)$$

where, for simplicity, we dropped the normalization constant (log(*N*)) in the definition of *H*. Consider the small perturbation of adding one feature to the matrix *A*. The assumption of a small

perturbation generally holds for a large enough number of features. Using equation (7), we can write the resulting change of H as

$$TdH = dK + (1 - \frac{K}{T})dT \qquad (12)$$

If an added feature projects only on the first PC, it can change only the first singular value. It follows then that

$$dT = dc_1, \;\; dK = -dc_1(1 + \log(c_1)) \qquad (13)$$

Plugging the terms in (9) into equation (8), we arrive at

$$TdH = \frac{TdK + (T-K)dT}{T} = -\frac{dc_1}{T}(K + T\log(c_1)) < 0 \qquad (14)$$

which means that adding such a feature always leads to reduction of entropy.
To complete the proof we show that the right hand side is indeed negative. T is positive, and so is the sum of the two terms in brackets, since c_1 is the leading eigenvalue and the following inequality holds:

$$-K = \sum_1^N c_j \ln(c_j) < T\log(c_1) \qquad (15)$$

We now prove that $dc_1>0$. Note that, by definition,

$$dc_i = \sum_{m,n} V_{mi} C_{mn} V_{ni} \qquad (16)$$

The first order perturbation of the eigenvalues of C is related to the change of the original matrix C by the original unitary transformation V. This follows from the unitarity constraint on V

$$\sum_m dV_{mi} V_{mi} = 0 \qquad (17)$$

and is the discrete analog of the Hellman-Feynman theorem [49], [50], [51].
Adding a row to A, i.e. adding the feature vector f^{M+1} of size N, the Gram matrix C changes to

$$C_{mn} \rightarrow C_{mn} + f_n^{M+1} f_m^{M+1} \qquad (18)$$

Plugging it back into equation (12), we conclude the proof with showing that dc_1 is positive according to:

$$dc_i = \left(f^{M+1} \cdot V^i \right)^2 \qquad (19)$$

where V_i is the i-th eigenvector of C.
Adjusting appropriately S and K, it is easy to prove this also for the sum of squares and the geometric mean functions mentioned in section 4.2.2.

4.7.2 When is UFF applicable?

We present two measures that allow for a separation between datasets on which UFF is effective, from those in which it is not. The first is SE, an entropy-like measure on normalized squares of UFF score-values.

$$w_k = \frac{Score_k^2}{\sum_{i=1}^{M} Score_i^2} \qquad (20)$$

$$SE = -\frac{1}{\log(M)} \sum_{k=1}^{M} w_k \log(w_k) \qquad (21)$$

and the second is VE, an entropy-like measure on the variance-values (i.e. variance of feature-values on all instances)

$$z_k = \frac{Var(f_k)}{\sum_{i=1}^{M} Var(f_i)} \qquad (22)$$

$$VE = -\frac{1}{\log(M)} \sum_{k=1}^{M} z_k \log(z_k) \qquad (23)$$

Suitable datasets can then be defined as those lying below certain thresholds in both measures. We tested more than a dozen 'suitable' and ten 'not-suitable' datasets (not shown) using UFF and clustering algorithms. It seems that combining the two measures using the geometric mean provides the best test for applicability. We found 'suitable' datasets to lie below a threshold of 0.8 of the combined score.

4.8 Supplementary Material

Tables S1-S22 of the supplementary material are found in http://adios.tau.ac.il/UFFizi/supp/ and on the attached CD.

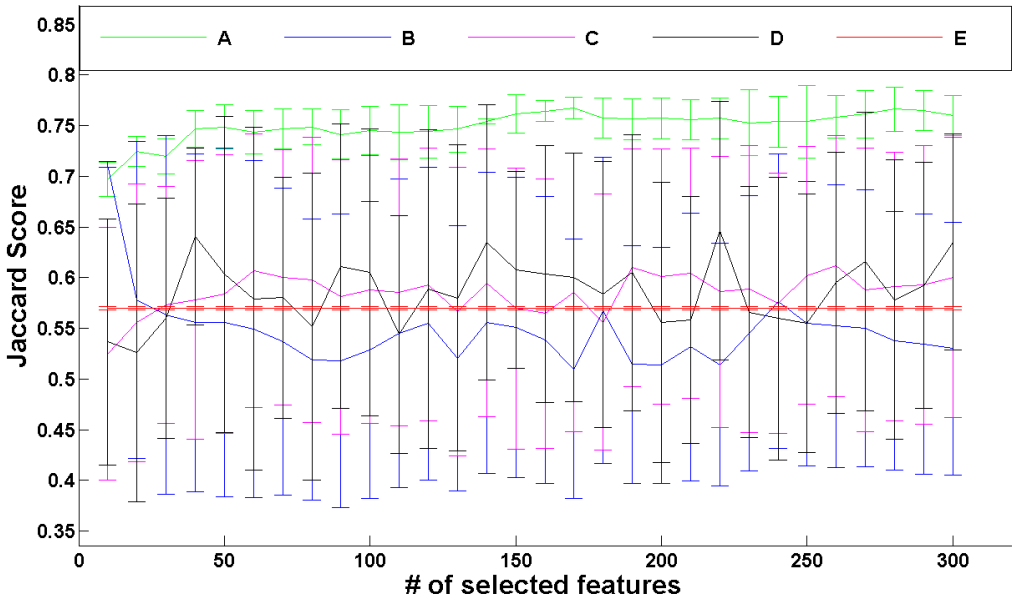


Figure S1. Comparison of UFF with other selection methods on the Melanoma dataset. Jaccard scores of clustering results for different selection methods on the melanoma dataset. Tested methods include (A) UFF, (B) Variance, (C) Feature entropy, (D) Random selection and (E) All features. Error bars denote standard deviation across different k-means runsClustering of 54 samples of GBM Agilent G4502A_07_1.4.2.0 array, colors and shapes denote different clusters. Image displays projection on principal components 2-4

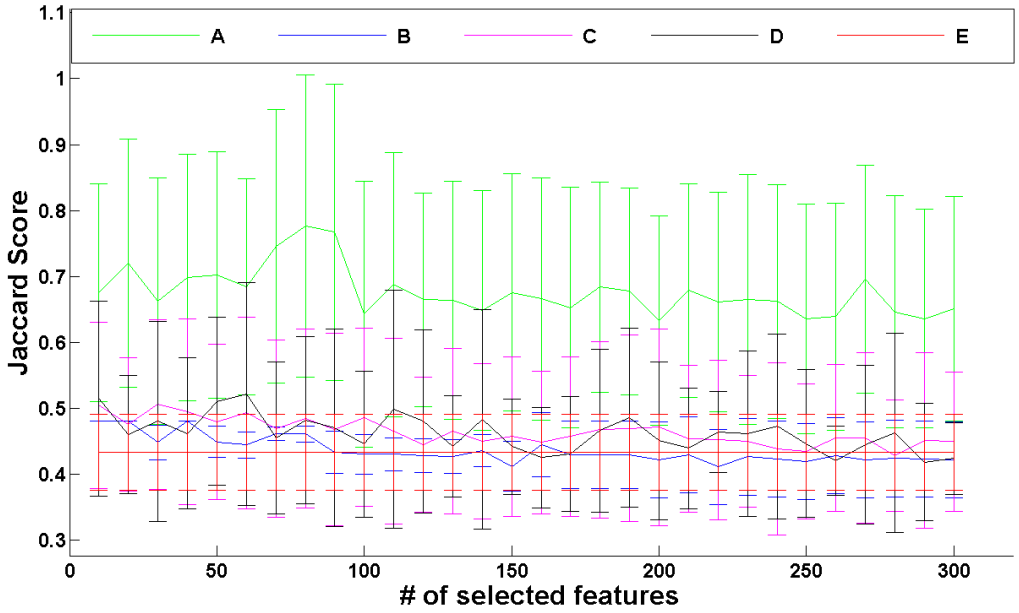


Figure S2. Comparison of UFF with other selection methods on the HIV dataset.
Jaccard scores of clustering results for different selection methods on the melanoma dataset. Tested methods include (A) UFF, (B) Variance, (C) Feature entropy, (D) Random selection and (E) All features. Error bars denote standard deviation across different k-means runs

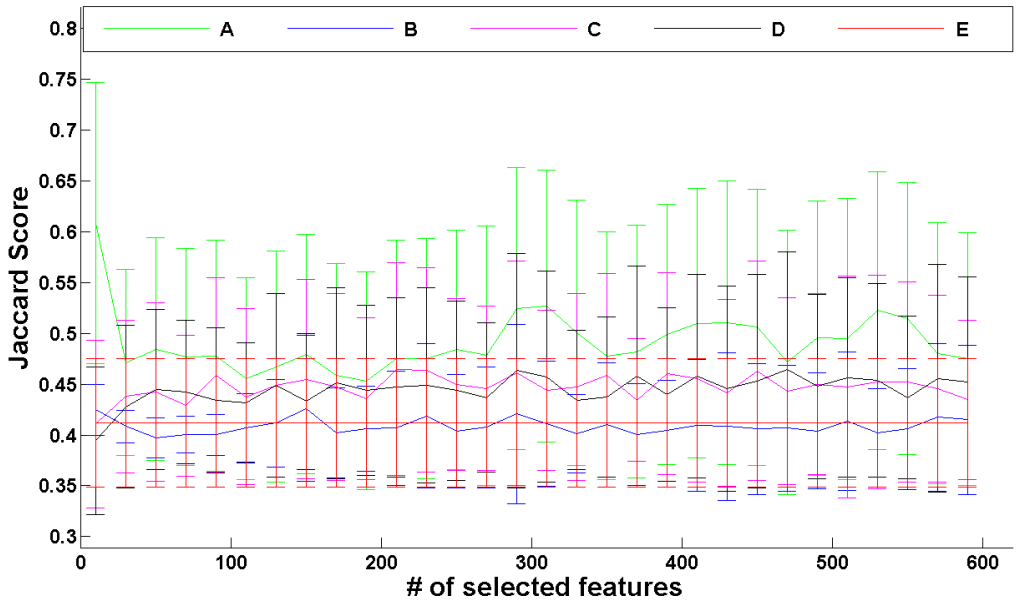


Figure S3. Comparison of UFF with other selection methods on the Hepatitis-C dataset.
Jaccard scores of clustering results for different selection methods on the melanoma dataset. Tested methods include (A) UFF, (B) Variance, (C) Feature entropy, (D) Random selection and (E) All features. Error bars denote standard deviation across different k-means runs

Part 2

Chapter 5

Extraction of Common Peptides (CPs)

5.1 Introduction

The analysis of protein sequences forms a valuable tool in protein function prediction. The primary method for sequence analysis is sequence similarity detection, typically implying homology, which may further imply structural and functional similarity. Many methods focus on pairwise or multiple sequence alignment [1], [2], [3], [4], [5]. Sequence alignment provides a distance metric that enables relating an un-annotated protein to a close annotated protein. Inter-protein distances may also be used for forming a vector of features describing the protein, which can then be exploited for the task of classifying them [6]. Other methods extract alternative features from protein sequences, including number count of different amino acids in the sequences (also termed AAC – Amino Acid Composition [7]) or using the physico-chemical properties of the amino acids [8, 9].

Another alternative to the standard sequence alignment is the identification of sequence motifs. Properly chosen motifs are expected to focus mainly on key regions in the protein and thus reduce noise from other regions. These motifs can span a feature space in which proteins may be represented and compared. Conventional motif extraction approaches construct motifs in terms of position-specific weight matrices, or use hidden Markov models and Bayesian networks, hence are supervised to some extent [10, 11].

MEX is a motif extraction algorithm that serves as the basic unit of ADIOS [13], an unsupervised method for extraction of syntax from linguistic corpora. MEX extracts motifs from sequence data of proteins in an unsupervised manner, without requiring over-representation of its amino-acid motifs in the data set. MEX motifs are deterministic strings in contradistinction to position-specific weight matrices or regular expressions. Based on MEX extracted motifs, [12] have introduced a method for classifying enzymes based on Specific Peptides (SPs).

In the SP method, motif extraction was carried out in an unsupervised fashion as a first step, followed by supervised selection from the resulting motifs according to their specificity to levels of the Enzyme Commission (EC) 4-level functional hierarchy.

The extraction of Common Peptides (CPs) utilizes MEX in a different manner. Instead of applying MEX to all sequences in an unsupervised manner, we apply MEX in a supervised fashion to

individual families of proteins, which may be families of enzymes belonging to certain EC numbers. Further processing is applied to the resulting set of motifs, including selection of motifs containing more than 4 amino-acids and elimination of degeneracy by removing motifs that contain other motifs. This defines a set of Common Peptides (CPs) characterizing the protein family. As opposed to the Specific Peptide methodology, there is no requirement that the motifs will not be found in other protein families in the training set. The distribution of CPs in the protein family, however, is easily distinguished from the distribution outside the protein family which highly resembles a random model. This is exemplified in section 5.1.1.

The protein family characterized by the set of CPs may be studied in several directions. The CPs constitute an inter-family conservation signal, often overlapping functional sites on the protein [14]. The first direction is to use the CPs to map important domains on the protein sequence which may have functional significance.

A second direction is to use search methodology in order to decide whether a queried protein belongs to the same family, on the basis of the CPs amino acid coverage of a given protein sequence. This task has a clear advantage over sequence similarity methods in the arising field of metagenomics, where only segments of DNA are provided, rendering the use of sequence alignment doubtful.

A third direction defines a feature space spanned by the CP list. Using this feature space, we reveal intra-family clusters, related to different functionality or evolutionary events during the development of the protein family. A final direction involves reconstruction of CPs on a given phylogenetic tree, tracking ancient genomic evolutionary events in the history of the protein family.

We present and example of ThyA and ThyX enzymes in section 5.1.1 to demonstrate the CP framework.

5.1.1 ThyA and ThyX: an example of CP methodology

ThyA is the classic thymidylate synthase family. Organisms that lack thyA possess an alternative unrelated enzyme, thyX, performing the same function. A small number of organisms possess both thyA and thyX. We have analyzed data [15] containing thyA sequences from 298 species and thyX sequences from 136 species. Only 13 species have both enzymes. ThyX exists almost exclusively in Bacteria, while thyA reside in all kingdoms.

MEX was applied to the thyA and thyX sequences, extracting for each type of enzyme its CPs. 313 and 168 distinct CPs were obtained for thyA and thyX respectively, covering 297 and 133 sequences of the two types of data, i.e. occurring at least once on more than 98% of the data. Species lacking

CPs may have very divergent sequences from all other species. An especially interesting case is that of Bacillus thuringiensis, containing two thyA enzymes and lacking CPs. Other species lacking CPs of thyX enzymes are *T. denitrificans*, *S. wolfei* and *M. thermoacetica*.

ThyA enzymes share a motif known as Prosite signature PS00091. This is a large motif, containing 8-13 non-specified amino-acids in the middle. CPs of the thyA enzymes are found to cover each of the two parts of the Prosite motif separately. ThyX enzymes share the motif RHRX7S [17]. The RHR prefix of this motif exists on seven of the CPs of thyX.

Figure 1 display an example of a thyX sequence of *D. Discoideum* and the list of CPs covering it. Nine CPs have hits on this sequence (shown in red). Two pairs of CPs are overlapping on this sequence. Each member of these pairs can be found without its overlapping companion on other sequences. The amino acid coverage of this sequence is 45 (the number of red characters in figure 1).

Found 9 hits

Peptide	Location
ARVSYG	67
DKGLI	81
LIRYL	84
ARQWIRHR	113
RHRTA	118
SARYS	128
YTEWYW	205
DLHNL	213
FLRLR	220

Mapping the peptides on the protein sequence

Red characters denote the location of the Peptide Matches

```
---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
MGLDIQTEIDKIVIEKVKPEVEYYDVMGGSHRWEVKVHDHGKVALVDTMPRLAPVGQTADFSICQAARVSYGAGTKKVTEDKGLIRYLYRHQHTSPFEMV
EFKFHCVMPVFIARQWIRHRTANVNEYSARYSVLPDKFYHPSIEEVRKQSTSNRQGGEEALEPKTAQEFLDYLDKVEENYKTYNELLEKGLSRELGRIGL
PVSIYTEWYWKIDLHNLFHFLRLRMDSHSQKEIRDYANTIFALIRPIVPVACEAFIDYAFESLKLTRLEIEAIRTGSPLNTTNKREIEEFEEKKKLLFPN
TQA
```

Figure 1. An example of a thyX sequence and the nine CPs covering it. The sequence is displayed in blue and the CPs hitting it are marked in red.

5.1.1.1 Coverage by CPs

We have studied the occurrence of CPs (number of hits) on enzyme sequences of the training set, and compared it to the occurrence of the same CPs on unrelated enzymes. Since CPs have not been selected according to specificity to a particular EC number, they may be found on sequences of enzymes whose function is unrelated to that of the family from which they were extracted. Nonetheless the occurrence distribution, as shown in figure 2, is very different. Figure 2 compares the distribution of thyX sequences, covered by various number of CPs. As displayed in figure 2, most of the thyX sequences have more than four CPs hitting them, where some have up to 31 CP

hits. In comparison, unrelated enzymes may have one CP hit, and rarely two hits. These numbers are consistent with a background random model, which randomly permutes the proteins and searches for matches of CPs on this permuted set. Within the family of proteins from which the CPs were extracted, one finds characteristically many CPs (average of 12 in the case of thyX) on the same sequence. Similar results are observed for thyA (not shown).

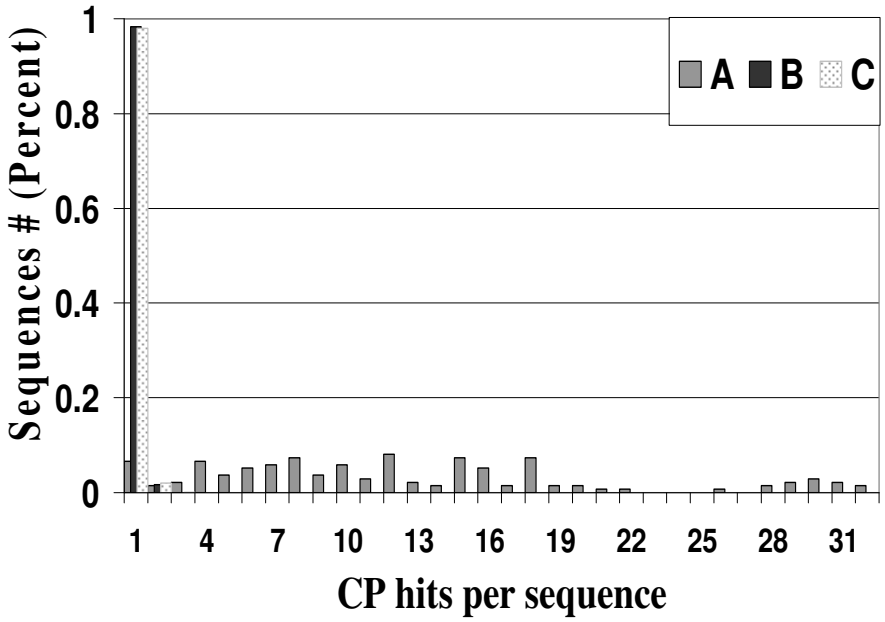


Figure 2. Number of CPs observed on each of the thyX sequences (A), is compared with the observation on sequences from all other enzymes (B), and with that of a random model (C). All three cases are normalized to total area = 1.

5.1.1.2 Biclustering of thyA and thyX

We provide here an example of the feature space spanned by the CP list. Applying biclustering to the matrix of species *vs* CPs of the thyX enzyme we obtain the results displayed in figure 3 (for explanation of the bi-clustering algorithm, see section 6.4.5). A clear biclustering pattern can be observed, with some CPs being intermediaries (i.e. connecting) between two or three clusters of organisms.

Next we apply the same procedure to the thyA data. The results, shown in figure 4, have completely different behavior: the clustering pattern of the thyX data is not observed in the thyA data, where

most species are contained in one large cluster. This may suggest that thyA evolved in a different way from thyX, e.g. thyA could have evolved from a single common ancestor protein, whereas thyX may have evolved from different origins. It is interesting to observe that the similarity of thyX sequences is much smaller than that of thyA ones (mean Smith-Waterman alignment e-value for thyA is 8.5e-6, while for thyX it is 0.007).

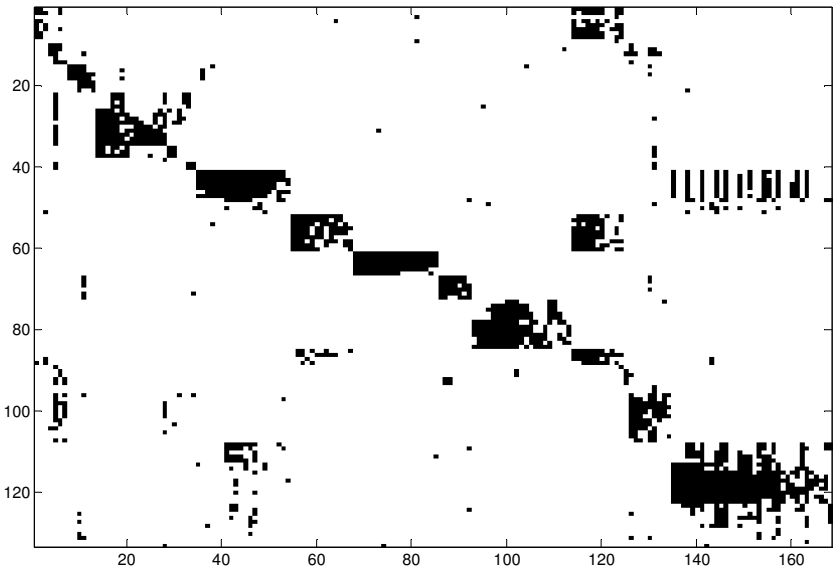


Figure 3. Biclustering of the matrix of species (rows) vs CPs (columns) of the thyX data

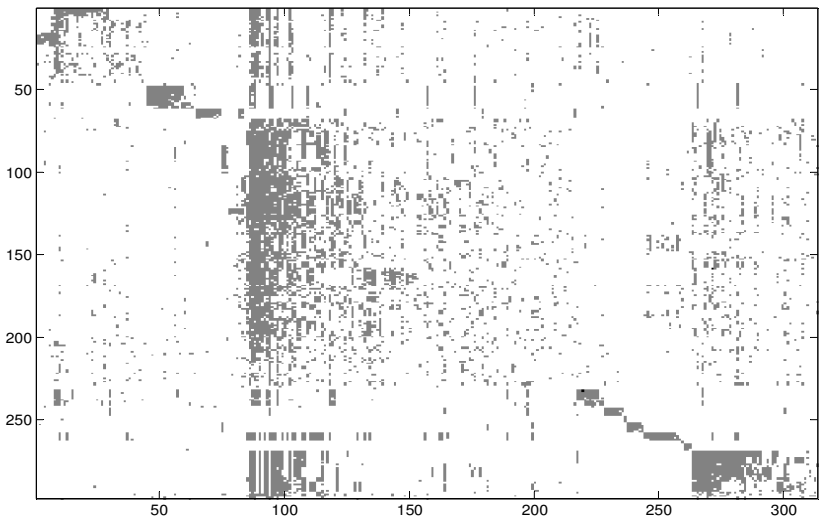


Figure 4. Biclustering of the matrix of (rows) vs CPs (columns) of the thyA data.

While thyX species form CP-disjoint clusters, thyA species fail to form such disjoint clusters. It is interesting to note that these statements hold also for Mycobacterium and Corynebacterium families that contain both thyA and thyX enzymes: while their CPs for thyX belong to a disjoint set, their thyA CPs are shared amongst multiple species (not shown).

Some of thyA and thyX species (72 of the former and 37 of the latter) appear in the tree of life (ToL) constructed in [16]. The tree of life is a tree connecting different species according to phylogenetic relationship of key genes that are common to all those represented. The CPs can be reconstructed on the tree and see if the inter-species relationships match that of the tree, based on sequence alignment. In addition, CPs connecting remote branches of the tree may point to lateral gene transfer (LGT) events.

The same biclustering algorithm can be applied to species containing thyA and thyX that appear on the ToL. We compared the clusters found for thyX sequences existing on the tree with the positions of their species on the ToL [16]. The results are displayed in Figure 5.

Most of the clusters correspond to species families or adjacent species on the ToL. There are three exceptions (clusters 1, 6 and 10) containing species which lie far apart on the ToL. The notably far species on cluster number 1 is D. Discoideum, the only Eukaryote known to contain thyX. The closeness in CP space suggests the occurrence of an LGT event between the Treponema family and D. Discoideum. This speculation is supported by the analysis of [15], who argued that D. Discoideum and Treponema subtree share a close ancestor. Another example is in cluster 10, where D. vulgaris, C. perfringens, G. sulfurreducens and B. cereus share CP space similarity, although far apart on the tree. This is also supported by [15], where they show homologous LGT between the Clostridia and delta-proteobacteria groups and proximity of all four species on their constructed phylogenetic tree.

Interesting results are also obtained on species containing thyA that appear on the ToL (figure not shown). While vertebrates cluster together, other eukaryotes appear in different clusters, sharing no or few CPs with the vertebrates (e.g. C. elegans, D. melanogaster and S. cerevisiae are on one disjoint cluster and O. sativa and A. thaliana on another).

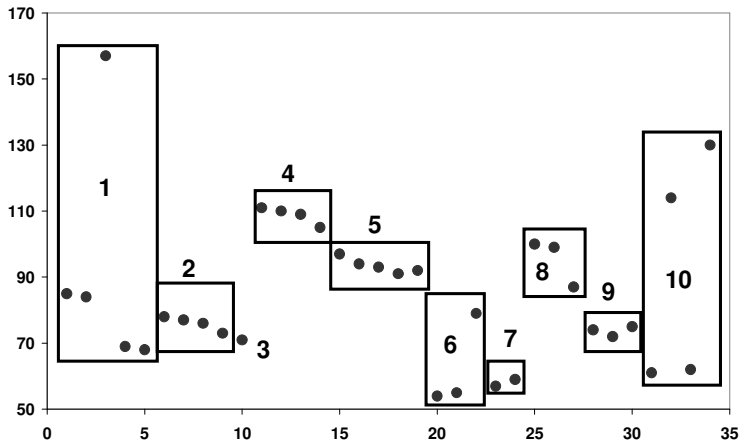


Figure 5. Location of the thyX species on the ToL (y-axis) as a function of the location calculated according to the bi-clustering algorithm. The analyzed species are a subset of the ones in figure 3, because many of the latter were not included in the ToL. Rectangles denote clusters.

5.1.2 References

1. Smith TF, Waterman MS: **Identification of common molecular subsequences**. *J Mol Biol* 1981, **147**(1):195-197.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
3. Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches**. *Science* 1985, **227**(4693):1435-1441.
4. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**(22):4673-4680.
5. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.
6. Liao L, Noble WS: **Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships**. *J Comput Biol* 2003, **10**(6):857-868.
7. des Jardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA: **Prediction of enzyme classification from protein sequence without the use of sequence similarity**. *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:92-99.
8. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ: **SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence**. *Nucleic Acids Res* 2003, **31**(13):3692-3697.
9. Cai CZ, Wang WL, Sun LZ, Chen YZ: **Protein function classification via support vector machine approach**. *Math Biosci* 2003, **185**(2):111-122.
10. Durbin R, Eddy SR, Krogh A, Mitchison G: **Biological sequence analysis: Probabilistic models of proteins and nucleic Acids Res**: Cambridge University Press.; 1998.
11. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.
12. Kunik V, Meroz Y, Solan Z, Sandbank B, Weingat U, Ruppin E, Horn D: **Functional representation of enzymes by specific peptides**. *PLOS Comp Biol* 2007, **3**(8):e167.
13. Solan Z, Horn D, Ruppin E, Edelman S: **Unsupervised learning of natural languages**. *in Proc Natl Acad Sci* 2005, **102**:11629-11634.
14. Meroz Y, Horn D: **Biological roles of specific peptides in enzymes.** *Proteins: Structure, Function, and Bioinformatics* 2008, **Online**.

15. Stern A, Mayrose I, Shaul S, Gophna U, Pupko T: **On the evolution of thymidine synthesis: a tale of two enzymes and a virus.** *Submitted for publication* 2008.
16. Ciccarelli FD, Doerks T, Mering Cv, Creevey CJ, Snel B, Bork P: **Toward Automatic Reconstruction of a Highly Resolved Tree of Life**. *Science* 2006, **311**(5765):1283 - 1287.
17. Leduc D, Graziani S, Lipowski G, Marchand C, Marechal PL, Liebl U, Myllykallio H: **Functional evidence for active site location of tetrameric thymidylate synthase X at the interphase of three monomers**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(19):7252-7257.

Chapter 6

Common peptides shed light on evolution of Olfactory Receptors[8]

6.1 Background

Odor recognition in vertebrates is mediated by a large superfamily of olfactory receptor (OR) genes, G-protein coupled receptors (GPCRs) with seven trans-membrane domains [1], [2]. Whole genome studies discovered hundreds of intact ORs in the vertebrate genome, ranging in size from ~100 in fishes to ~1000 in mouse [3-6].

A recent study of OR evolutionary dynamics indicated the existence of nine ancestral genes common to fish and tetrapods, of which only two are found in birds and mammals. Specifically one of these, known as Class II, has expanded enormously in mammals [7]. Several studies have applied computational sequence analysis and phylogeny methods to study the evolution of the OR repertoire in vertebrates [7, 8]. One of these studies [9] used motifs to analyze human and mouse OR repertoires, focusing on classification of the motifs into classes and classification of the ORs using these motifs as features.

We adopt a different motif-based approach that extracts deterministic motifs, i.e. peptides, and explores their appearance along OR evolution. We apply the motif extraction algorithm MEX [10], the efficacy of which has been previously demonstrated in the study of enzymes [11] , to 4027 OR sequences of 10 vertebrates. A short explanation of MEX is also provided in the Methods section. The union of all motifs leads to a list of 2717 MEX-derived peptides, to be referred to as Common Peptides (CPs). These motifs can be mapped onto specific locations on the seven trans-membrane domains.

Following CP occurrences on ORs of different species we can trace the development of these domains with evolution. Using the Tree of Life, we perform an ancestral reconstruction of CPs and determine their evolutionary ages.

For each species we perform biclustering of the matrix of CP occurrences on ORs. Choosing CP groups according to their evolutionary age we get different clustering patterns.

The use of CPs for studying OR sequences enables us to explore different aspects regarding OR evolution than those uncovered by phylogenetic methods. It also enables us to uncover some fine

[8] Based on the paper *Common peptides shed light on evolution of Olfactory Receptors,* Assaf Gottlieb, Tsviya Olender, Doron Lancet and David Horn, BMC Evolutionary Biology 2009, 9:91.

details of OR groups that were previously studied using regular-expression motifs, due to the deterministic nature of our motifs (see also [12]).

6.2 Results

6.2.1 CP mapping on the Tree of Life

We used 4027 OR sequences representing the complete intact OR repertoires in 10 vertebrates (Table 1). We extracted a list of CPs by applying MEX to OR sequences of each species individually, followed by a unification procedure to remove redundancy (see Methods for a detailed description).

All CPs are tested for their occurrence on all ORs, irrespective of which species lead to their extraction. We define *species-specific CPs* as CPs observed only in one species.

On average an OR is matched by 48 CPs, covering 147 amino acids on its sequence. Some CPs partially overlap with one another. The total number of CPs found in sequences of one species (column 3 in Table 1) is highly correlated (Pearson correlation = 0.9) with the number of ORs per species (column 2 in Table 1).

Table 1. Distribution of 3983 OR sequences, total CPs and species-specific CPs according to species

Species	Number of ORs	Number of observed CPs	Number of species-specific CPs	Percentage of species-specific CPs
Pufferfish	44	193	11	5.7%
Zebrafish	97	352	60	17.0%
Frog	409	1179	143	12.1%
Lizard	120	945	17	1.8%
Chicken	78	644	15	2.3%
Platypus	250	1406	26	1.8%
Opossum	846	2030	48	2.4%
Dog	814	2083	40	1.9%
Mouse	978	2179	66	3.0%
Human	391	1889	8	0.4%

The percentage of species-specific CPs is particularly high in fish and frog (although less than 6% of the pufferfish CPs are pufferfish-specific, the percentage of fish-specific, including both fish, is 18%). The percentage of species-specific CPs drops significantly to an average of 2% in other species, with human having the smallest amount of species-specific CPs. This finding might be attributed to the difference between aquatic environment, characteristic of fish and the amphibian frog *X. tropicalis* that remains aquatic also in its adult life (see [13] and [14]), and terrestrial environments characteristic of the other species: presumably CPs were lost - together with their ORs (groups δ, ε, ζ and η in [7])– in terrestrial species that have developed later.

We evaluate the emergence and loss of CPs on a commonly accepted tree of life representation (figure 1), using the parsimony method (see details on the chosen method and other tested ancestral reconstruction methods in the Methods section).

We identify "*novel CPs*" as those that exist in the current ancestor/species but did not exist in previous ancestors, and "*lost CPs*" as those that do not exist in the current ancestor/species but did exist in the previous ancestor. CPs that date back to previous ancestors are referred to as "*conserved CPs"*.

The analysis detects one major addition of novel CPs in the ancestor of tetrapods, A2. Judging by [15] the branch length between A1 and A3 is about the same as that between A3 to A6. 47% of the CPs at A6 are novel with regard to A3. This should be compared with the fact that 75% of CPs at A3 are novel with regard to A1. We thus may conclude that the main expansion of OR CPs has taken place at, or before, A3.

Reptiles have suffered major losses of CPs, a trend that was further increased in chicken. Another major loss occurred in pufferfish.

Interestingly, while humans lost more than half of their ORs relative to other mammals, they lost only 11% of the CPs existing in A6. This suggests that some redundancy in mammalian ORs has been removed by OR pseudogenization in human. This result is surprising considering the fact that the human intact OR repertoire contains much less subfamilies relative to other mammals (according to HORDE classification system [16]). For example, there are 242 and 227 subfamilies in mouse and dog respectively, but only 175 subfamilies in human. Investigating subfamilies of mouse and

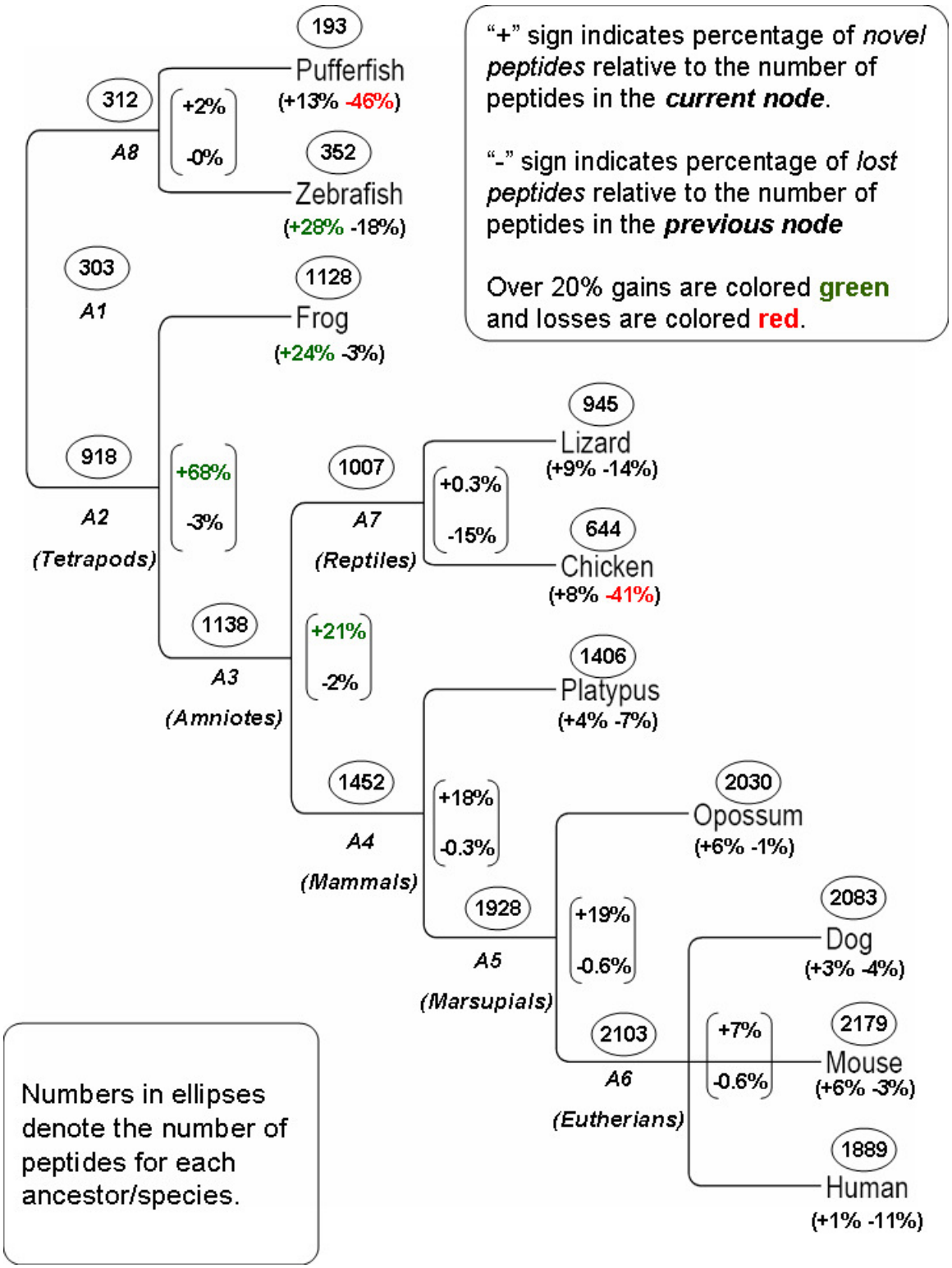


Figure 1. CP reconstruction on the tree of life. Number of CPs occurring in each species and parsimoniously estimated number of CPs occurring in each ancestor (in ellipses). Numbers in brackets indicate the percentage of *novel CPs* relative to the total number of CPs in the current node (+ sign) and the percentage of *lost CPs* relative to the total number of CPs in the previous node (- sign). Over 20% gains are colored green and lost are colored red. Ancestor names are enumerated from the most recent ancestor of fish and tetrapods (A1) to pufferfish and zebrafish ancestor (A8). As an example, zebrafish contains 97 novel CPs, which constitute 28% out of its 352 CPs. It also lost 57 CPs, which occurred in its ancestor, which constitute 18% of the CPs existing in A8.

dog ORs that are not matched by human subfamilies, we nonetheless find many of their CPs (68% of mouse CPs and 35% of dog CPs) elsewhere in. other human subfamilies. In other words, according to the CP perspective the similarity between human and mouse or dog is larger than observed by the sequence similarity which is the basis of the subfamily classifications. [17] hypothesize that the reduced sense of smell in human could correlate with the loss of functional genes. The high co-occurrences of CPs in functional human, mouse and dog genes hints, however, that the reduction of the human OR repertoire may not necessarily cause loss of functionality.

6.2.2 CPs that make a difference

The CP method extracts CPs that bear statistical significance. It is reasonable to assume that some of them also have biological significance. We first looked for CPs that differentiate between water-dwelling species (i.e. pufferfish, zebrafish and possibly frog) and purely terrestrial species. We find 10 CPs that exist in fish (one of them occurs also in frog) but not in any other land-dwelling species. Similarly, we find 44 CPs which are terrestrial specific (none of them exist in frog). Of special interest are CPs that reside in the outer region of the membrane (extracellular loops and the external half of the transmembrane domains). Such CPs might participate in ligand binding. Table 2 lists the CPs residing only in water-dwelling species. CPs that potentially play part in ligand binding are marked. Of particular interest is the CP "RLPLCG", which resides on the extracellular loop 2 and contains a Cysteine, possibly crosslinking with another Cysteine on the ORs.

Table 3 lists the CPs residing only in terrestrial species. CPs that potentially play part in ligand binding are marked. More than 2/3 of these CPs occur in ORs that belong predominantly (more than 40% of the total OR occurrences) to one HORDE family.

Table 2. CPs specific to water-dwelling species. CPs facing the extracellular side of the membrane are in bold.

CP	Domain	# of occurrences
RYILF	TM2	15
YGATGFYP	**TM2**	**6**
AGFFPR	**TM2**	**11**
LAYDRL	IL2	9
YHSVM	IL2	10
RLPLCG *	**EL2**	**17**
KFMQTC	IL3	8
ALKTC	IL3	16
QTCVPH	IL3	16
PPILNPL	TM7	13

Domains start from the N-terminal (N), through Transmembrane domains 1-7 (TM1-TM7), Intracellular loops (IL1-IL3) and extracellular loops (EL1-EL3) and end in the C-terminal (C)

* - appears also in frog

Table 3. CPs specific to land-dwelling species. CPs facing the extracellular side of the membrane are in bold.

CP	Domain	# of occurrences
NHTTV	N	30
QVLLF	TM1	53
TLMGN	TM1	89
GNLGM	TM1	211
LGNGTIL	TM1	20
NLGMI	TM1	181
FLSSLS	TM2	53
VDICF	TM2	71
CFSSV	**TM2**	**59**
GVTEF	**TM2**	**55**
TVPKS	**TM2**	**39**
TTTVP	**TM2**	**64**
PKMIAD	**TM2**	**19**
MLVNF	**TM2**	**153**
LPRML	**TM2**	**39**
KVISF	**EL1**	**85**
ISFTGC	**EL1**	**45**
GCATQ	**TM3**	**117**
SYSGC	**TM3**	**47**
AQLFF	**TM3**	**107**
LVAMA	TM3	122
NPLLY	IL2	349
PLHYL	IL2	110
PLLYP	TM4	68
SWLGG	TM4	54
GLFVA	**EL2**	**60**
YTVIL	**TM5**	**50**
SYGLI	TM5	34
LAVVTL	TM5	23
ILRIR	IL3	142
LRIRS	IL3	159
RKALS	IL3	161
LLFMY	TM6	61

LFFGP	TM6	133
AYLKP	**TM6**	**54**
TYIRP	**TM6**	**29**
YLRPSS	**TM6**	**50**
IYARP	**TM6**	**49**
VALFY	**TM6**	**50**
RPSSS	**TM6**	**86**
LFYTI	**TM7**	**115**
EVKGA	C	108
GALRR	C	65
AMRKL	C	61

Domains are the same as in table 2.

6.2.3 GPCR remote homologies

ORs are part of a larger protein superfamily of G-Protein Coupled Receptors (GPCRs). We searched 967 chicken, human and mouse non-OR GPCRs taken from [18] and [19] and found 526 of the OR CPs to appear in this dataset (figure 2). The number of CP occurrences (hits) on an OR is easily distinguishable from other GPCRs. The number of CP hits on non-OR GPCRs exceeds that of a random model, from which one expects to observe at most one or two CP hits. Our observation of up to 6 CP hits for some non-OR GPCRs indicates an ancestral relation between ORs and some non-OR GPCRs, i.e. remote homology (see histograms S6-S9 in Additional file 1] and explanation of the random model in the Methods section).

Figures S1 and S2 are histograms of the same kind for chicken and mouse respectively.

In figures S3-S5 we study the loci of OR CPs on non-OR GPCRs in chicken and mammals respectively. Sharp peaks in mammals correspond to known motifs [20]. No sharp peaks are observed in chicken.

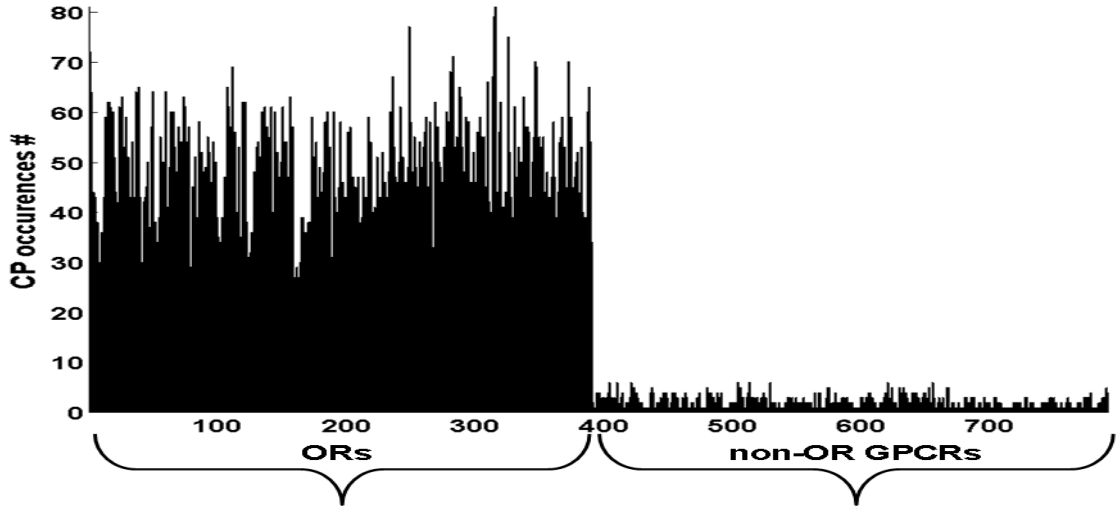


Figure 2. CP occurrences on human GPCRs. The number of CP occurrences (hits) for each of the 391 human ORs (ordered by HORDE) and, followed by 400 human non-OR GPCRs (ordered by [14]).

6.2.4 Locations of CPs on the OR sequence

We investigate the locations of the CPs along the 7 trans-membrane (TM) domains. The resulting histograms are compared with conservation loci of single amino-acids [21]. Locations are determined relative to a highly curated multiple alignment of human and mouse ORs. The histogram in figure 3 displays the relative coverage by CPs of each position along the OR chain (see Methods section 3.4 for a description of normalization of positions between ORs). Highly conserved positions of amino-acids, as deduced by [21] from mouse and dog data, are indicated by red coloring of the histogram on 65 positions.

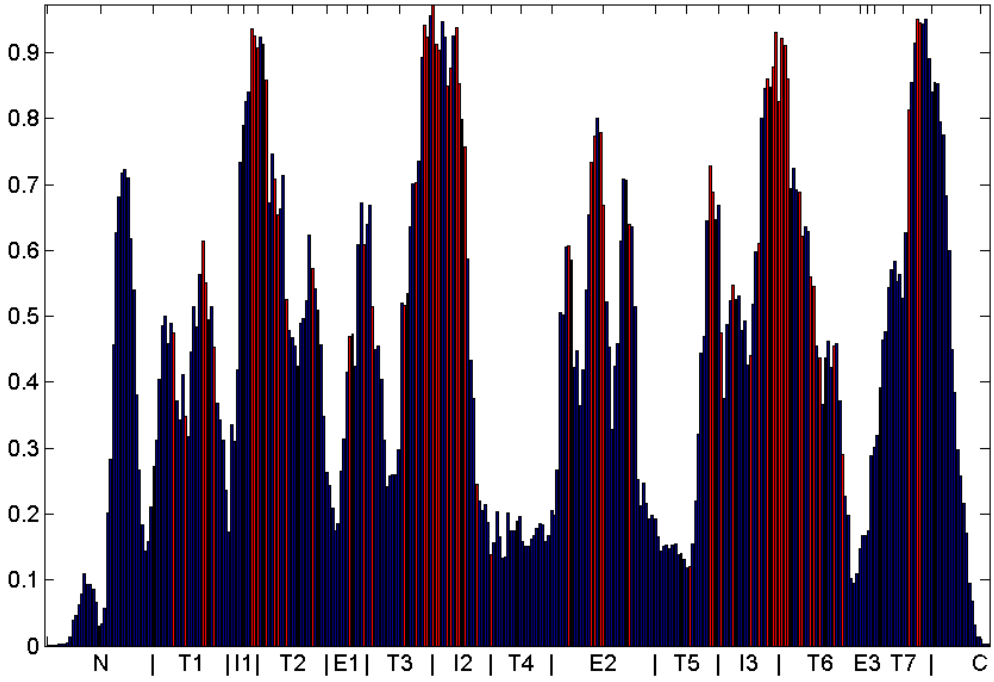


Figure 3. CP coverage of positions along the OR sequence. Positions start from the N-terminal (N), through Transmembrane domains 1-7 (T1-T7), Intracellular loops (I1-I3) and extracellular loops (E1-E3) and end in the C-terminal (C). 65 known highly-conserved positions are indicated by red.

Figure 4 shows the CP position coverage for four species. Figures displaying all CP positions for these three species, all other species, assessed ancestor CPs, novel and lost CPs, are provided in (figures S10-S15) [see Additional file 1].

Figure 4 indicates four regions which are highly populated with CPs along all vertebrate evolution. These regions are marked using a threshold drawn at 60% sequence population in zebrafish, displayed in figure 4B. All four regions reside in the interface between the transmembrane domains and the intracellular regions (IL1-3 and the C-terminal). These regions may be connected to structural constraints in the interface that binds the G-proteins. Figures displaying OR coverage by position for all other species ranging from frog to human look very similar (figures S10, S11 [see Additional file 1]). We observe that CPs within some regions have developed much higher coverage only in tetrapods. These regions are marked in figure 4D. They are: the end of the N-terminal, the interface between extracellular loop 1 (EL1) and TM1 and TM2 and the middle of extracellular loop 2 (EL2). Most of the newly emerged regions are facing the extracellular side of the membrane. This imposes structural constraints on the regions connected to odorant binding and might be specific to airborne odorants.

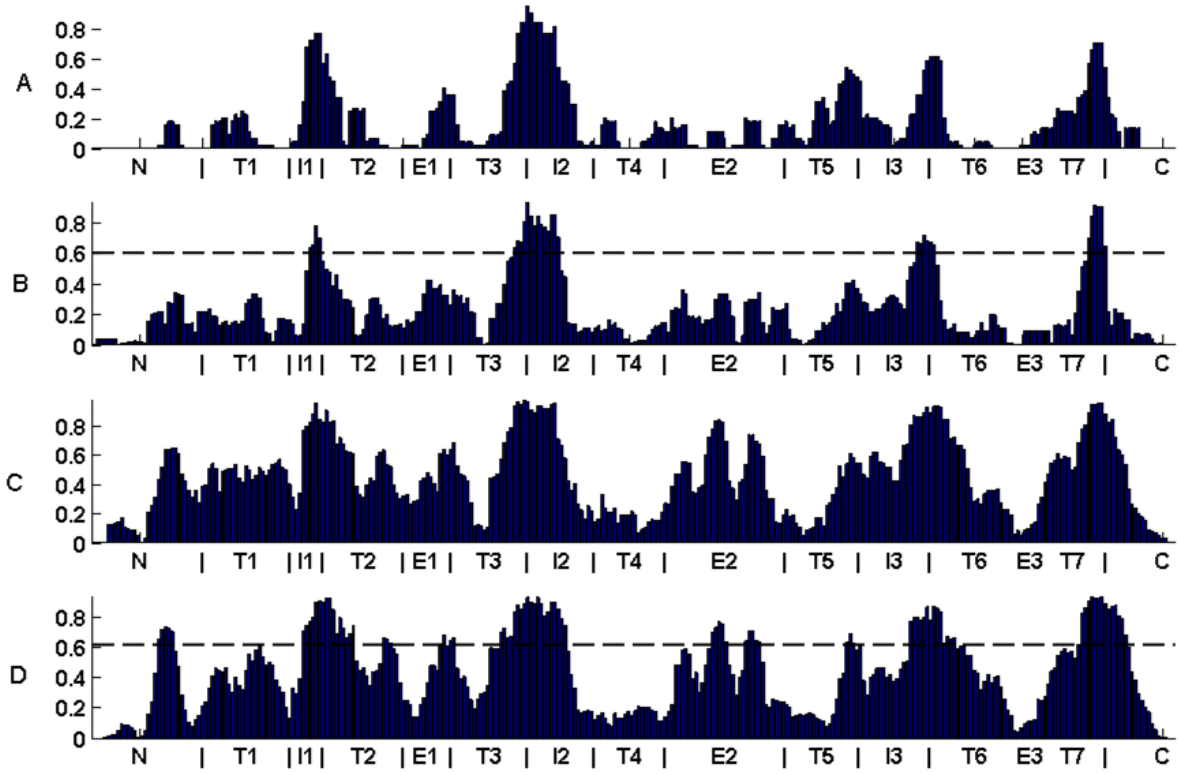


Figure 4. CP coverage of positions along the OR sequence for selected species. CPs coverage of positions along the OR sequence for pufferfish (A), zebrafish (B), Frog (C) and Human (D). Thresholds mark the regions that are common to all ten species (B) and new to vertebrates (D). Positions are the same as in Figure 3.

6.2.5 CP-space reveals internal clusters

Using biclustering, we obtain simultaneous co-occurrences of ORs and CPs for each species. This provides a powerful visualization and allows the study of evolutionary trends across species. Details of the biclustering algorithm and its application are found in the Methods section.

We perform the analysis using different sets of CPs characterized by their evolutionary ages.

First, we apply the procedure to zebrafish ORs, represented either by the *conserved CPs*, i.e. CPs shared with tetrapods (A1) or by zebrafish novel CPs (see figure 1 for reference). There are only nine CPs novel to A8 (the common ancestor of zebrafish and pufferfish) hence they are not used in

the clustering analysis. The results are displayed in figure 5. We identify an interesting pattern in this figure. Zebrafish novel CPs form almost disjoint biclusters, while OR clusters based on conserved CPs (CPs originating high in the tree) tend to share CPs (figure. 5A). Conserved CPs cover almost all ORs (seven ORs did not pass the threshold of minimal CP number specified in the Methods section). Novel CPs cover only half of the ORs.

We identify ten clusters in zebrafish using ancestral (A1) CPs and six using zebrafish-novel CPs. Each of the latter six clusters matches one of the former clusters. The detailed cluster assignments are displayed in the supplementary material [see Additional file 1].

Novel CPs emerge from speciation and duplication events occurring after the split of fish from A1. We find 10 ORs that do not have any novel CPs in zebrafish and fish common ancestor (A8). This can serve as a first estimate of the number of ORs that existed in A1. They reside in the OR clusters indicated by red circles in Figure 5A.

Classification of zebrafish ORs into groups has been studied by [7] and [22]. Both found eight groups with different OR membership (four groups of [7] and one of [22] contain only one OR each). Biclusters of novel CPs (Figure 5B) map perfectly to some groups (groups δ, ζ and η of [7]), where some groups are further split to reveal finer details (e.g. groups δ and ζ of [7] and group E of [22] are split into two biclusters). The 10 ORs which contain no novel CPs have members only from groups δ, θ and κ of [7]. For mapping between our clusters, and the groups of [7] and [22], see additional files 2, 3 and 4.

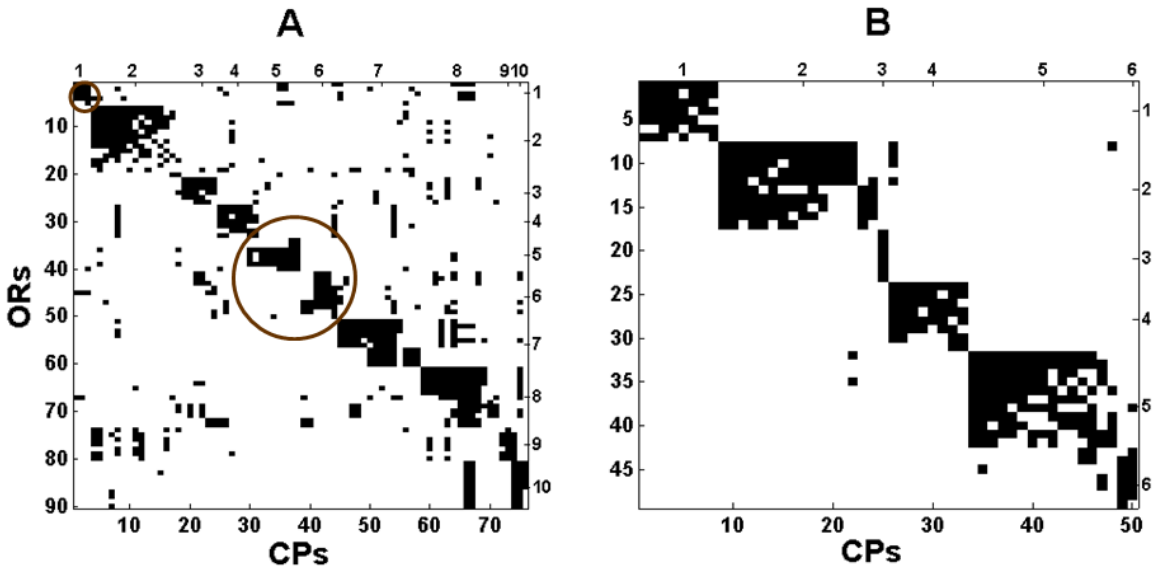


Figure 5. Biclustering results of Zebrafish. Y-axis corresponds to ORs and X-axis to (A) A1 (root ancestor) CPs and (B) zebrafish novel CPs. Circled clusters in (A) have no corresponding biclusters of novel CPs in B.

The biclustering algorithm allows us also to differentiate between the different zebrafish clusters. The assumption is that OR clusters which relate to recent ancestry might also bear functional similarity. While some of the CPs that differentiate between the OR clusters are conserved remnants of duplication events, other CPs represent segments of these ORs that might contribute to a common functionality of the OR cluster. A table of the CPs of each cluster is provided [see Additional file 5].

Pufferfish has few novel CPs. Biclusters formed using CPs belonging to A1 look similar to the ones displayed in Figure 5A. The biclustering of pufferfish appears in figure S16 [see Additional file 1].

Figure 6 displays biclustering results of frog. Three sets of CPs are being used, those novel to A1, novel to the tetrapods' ancestor (A2) and novel to frog. Ancestral CPs form noisy clusters, while CPs novel to frog form almost disjoint clusters, similar to the zebrafish biclusters. As in zebrafish, the number of ORs covered by CPs drops with the age of the CP (i.e. the node in the ToL where it first appears). We identify nine clusters using CPs novel to frog. They map almost perfectly to clusters identified using either novel CPs of A1 or A2 [see Additional file 3].

Unlike zebrafish clusters, not all the A1 and A2 conserved CPs form identifiable biclusters. This suggests that they have been subjected to a higher mutation rate than observed in zebrafish, which may relate to the appearance of class II ORs in frog [23]. The clusters in figure 6c relate to the groups γ and δ of [7], [see Additional file 4].

Chicken and lizard have too few novel A3 and A7 CPs, to construct biclusters. The novel CPs of chicken form one big cluster, while novel CPs of lizard form small disjoint clusters. Novel CPs to A1 and A2 also show difference between chicken and lizard. While the former reveals a robust big cluster, the latter show no clusters at all. This implies large number of recent duplications in chicken. The biclustering of chicken and lizard appear in figures S17-S18 [see Additional file 1].

Biclusters in mammals are displayed in figures S19-S23 [see Additional file 1]. Biclusters are significant for CPs novel to A3- A6. They can be mapped to class I (fish-like) and class II (mammals-like) ORs, and to families of the Human Olfactory Receptor Data Explorer (HORDE). The mapping appears in Additional files 6, 7, 8, 9, 10, 11 and 12.

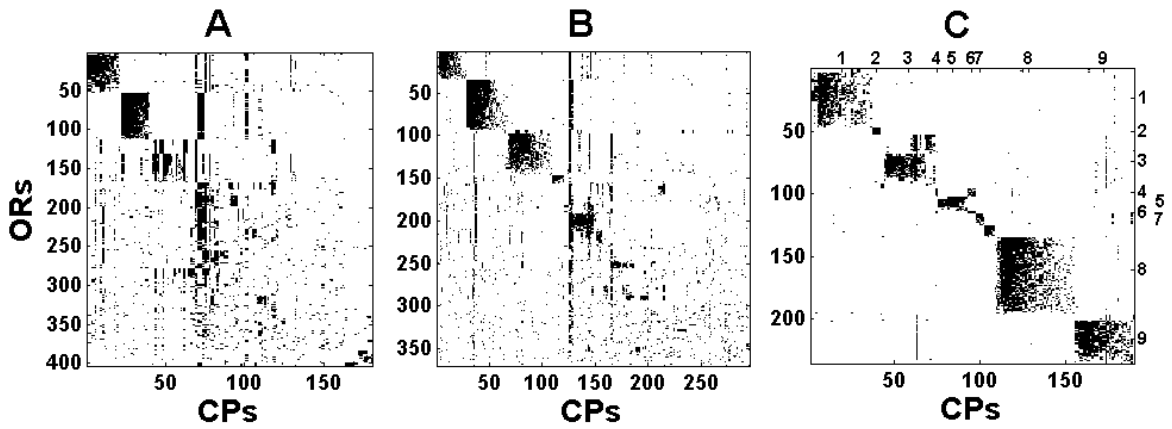


Figure 6. Biclustering results of Frog. Y-axis corresponds to ORs and X-axis to CPs novel to A1 (A), to A2 (B) and CPs novel to frog (C).

6.2.6 Novel CPs and mammalian families

Figure 7 shows the correspondence between mammalian CPs and the classification of the OR superfamily into families, using the HORDE classification system [16]. Class II (families 1-13) ORs contain predominantly CPs of A2. In contrast, class I (families 51, 52 and 56) ORs have equal distribution of novel CPs from A1 and A2. We also observe that family 3 almost ceased to evolve after A2 and families 9 and 11 stopped evolving after A3.

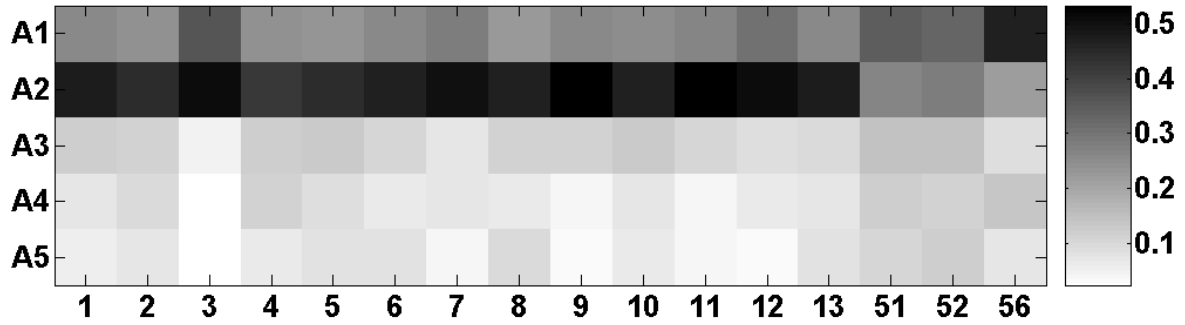


Figure 7. Distribution of CP age, novel to A1- A5 ancestors for each mammalian HORDE family. X-axis corresponds to family number. Color scale corresponds to percentage from the total number of CPs of each family, ranging from 0 (white) to 1 (black).

6.3 Discussion & Conclusions

We use CPs extracted by MEX (Motif Extraction algorithm) to study evolutionary processes in olfactory receptors. Such conserved CPs are known to have biological importance [24] and are expected to play structural and functional roles in olfactory receptors. Having extracted such CPs

from ten species, we use evolutionary constraints to further employ the extracted CPs in making sense of the complex relationships of ORs of different species with one another.

The evolutionary perspective is obtained by applying the parsimony principle to a tree-of-life accommodating the studied species. It allows us to construct an ancestral phyletic pattern of the presence or absence of CPs in internal nodes of the tree. Using this construction, we show that the number of species-specific CPs is relatively high in fish and frog, but remains fixed in terrestrial species. The species-specific CPs in the aquatic species might be related to ORs detecting water-soluble odorants. We observe a major emergence of CPs in the ancestor of tetrapods and major losses of CPs in pufferfish and in chicken. A surprising result stemming from this mapping is that although humans lost half of the intact mammalian ORs, they lost only 11% of the conserved CPs, suggesting a controlled process of loss of redundant ORs. In other words, the potential odorant recognition of humans may have suffered only a minor damage by the severe diminution of their OR repertoire.

CPs that differentiate between water-dwelling species and terrestrial species have potential biological significance and are candidates for further biochemical studies.

We show that some of the OR-extracted CPs exist in the general GPCR population, demonstrating the ancient origin of ORs and several other GPCRs.

The fact that the OR history stretches back to fish was made by [7] who claimed that 85%-90% of frog, chicken, mouse and human OR repertoires was constructed from duplication of a single fish OR of group γ, Dr3OR5.4. One or more of these 35 fish group γ CPs are also observed in 98% of the tetrapod ORs. This is larger than the coverage observed for CPs in any other fish ORs. These 35 CPs are also almost exclusively located in the five most conserved positions in figure 3 (boundary between IL1 and TM2, boundary between IL2 and TM3, middle of EL2, boundary between IL3 and TM6 and TM7). We point out, however, that major changes have occurred in other nodes of evolutionary history. By studying loci of CPs we identify two regions that show high CP coverage starting from tetrapods: the N-terminal and the middle of the second extracellular loop. This might imply that these regions are important for the adaptation of ORs to airborne odorants.

Gene multiplication events are most naturally exhibited by the existence of clusters of ORs. Using the evolutionary separation into novel and conserved CPs, we are able to demonstrate clean OR clusters. This is done by applying a biclustering algorithm to matrices associating CPs with ORs within species: clean clusters emerge when novel CPs are being employed. Results vary with increasing evolutionary age of the species in question. Our biclustering results of the species studied by [7], [22] (zebrafish, frog and chicken) generally support their phylogenetic models, but provide finer OR grouping and a cleaner selection of the responsible ancestor (where CP formation has

occurred). Finally, we are able to use the CP analysis to provide developmental details of OR families of the Human Olfactory Receptor Data Explorer (HORDE).

6.4 Methods

6.4.1 Data

For the described study we selected a set of 4027 intact olfactory receptors (ORs) from ten vertebrate species including pufferfish (*Takifugu rubripes*), zebrafish (*Danio rerio*), frog (*Xenopus tropicalis*), chicken (*Gallus gallus*), lizard (*Anolis carolinensis*), platypus (*Ornithorhynchus anatinus*), opossum (*Monodelphis domestica*), dog (*Canis familiaris*), mouse (*Mus musculus*) and human (*Homo sapiens*).

All mammalian, chicken and lizard OR sequences are available at the HORDE [16]. OR sequences of fish and frog were taken from the study of [7]. Lizard and Platypus ORs appear in [25]. The number of ORs for each species is listed in Table 1.

967 chicken, human and mouse non-OR GPCRs were taken from [18] and [19].

6.4.2 MEX algorithm

MEX is a motif extraction algorithm introduced by [10] as part of a method for grammar induction from texts and was later used on proteins [11]. Given a set of proteins, they are represented as different paths over a graph that consists of 20 vertices, corresponding to the 'alphabet' of 20 amino-acids. MEX proceeds by looking for convergence of many paths onto strings of amino-acids, and the subsequent divergence from such strings. The latter are defined as motifs if both convergence and divergence obey some statistical conditions. These conditions are imposed on context-dependent variable-order Markov chains that are constructed out of the data-paths. The algorithm has two parameters, η and α, specifying the amount of convergence/divergence and its statistical significance given the number of paths involved in the process. More information can be found on the website [26].

In the present analysis we ran MEX on the proteins of each species separately, using the parameter values η=0.9 and α=0.01. We restricted ourselves to peptides of length 5 amino-acids or more and appearing in at least 4 ORs. These peptides were merged into one list, where duplicates and peptides containing other peptides were removed. The resulting non-redundant list contains 2717 Common Peptides (CPs). Each of the CPs was then searched on the ORs of all species. CPs that appear only in the ORs of one of the studied species are defined to be *species-specific*.

6.4.3 Fitting CPs to the tree of life and phylogenetic analysis

We used the tree of life web project, available at [27] to construct the relationships between the species. The relations between the species is consistent with the tree of life of [15] . Dog, Mouse and Human were put under one common ancestor according to the tree of life web project, although there are other possible ancestral orders based on different set of genes (see also[28], [29]-[30]). Trying other arrangements for Dog, Mouse and Human did not alter the derived conclusions. The assessment of CP origins uses the Wagner parsimony, as implemented by the Phylogeny Inference Package computer programs PHYLIP. Similar results are also obtained by Dollo parsimony.

Since some CPs differ by only one amino acid from others, we have also checked whether loss and gain of a CP on any internal node corresponds to a mutation of a single amino-acid (interpreted as a loss of the CP) into another amino-acid (interpreted as a gain of a CP). We have found that the number of such events is negligible (1 such event in an ancestral node on average and 7 on average in the species, occurring mainly in chicken and lizard).

Following Parsimony estimation, each internal node A1-A8, and each species, has a list of CPs associated with it. We identify "*novel CPs*" as those that exist in the current ancestor/species but did not exist in previous ancestors and "*lost CPs*" are defined as those that exist in the current ancestor/species but did exist in the previous ancestor. CPs that date back to previous ancestors are referred to as "*conserved CPs"*.

6.4.4 Normalizing CP positions

Each CP contains a set of positions relative to the start of each OR. Due to variable N-Terminal length and gaps, we needed to normalize the different positions of each CP appearing in different ORs. We normalized the OR relative positions using ClustalW2 (available at [31]). We first aligned the five sequences used in [32] to construct a profile (replacing MOR257-1 that was not available in our set with MOR257-10). Each OR was then aligned to this profile.

6.4.5 Biclustering

Biclustering is performed on the ORs of each species, using subsets of CPs, each subset corresponding to a different origin on the tree of life. Each OR is represented by a binary vector that signifies the existence or non-existence of each of the CPs on its sequence. In order to clear noise, we first removed all ORs having less than 5 CPs from the relevant tree of life node. We then removed CPs that appear in less than 5 ORs from the remaining set. ORs left with no CPs after the previous removal were also removed. We used a bipartite spectral graph partitioning algorithm of [33]. Initially designed for documents and words, this bi-clustering algorithm handles sparse data

well. This algorithm produces biclusters of ORs and CPs. We augmented the algorithm to produce good biclusters' images. This was achieved by applying single linkage hierarchical algorithm for each produced bicluster and sorting each bicluster according to the hierarchical clustering, thus handling less homogenous clusters better. This augmentation of the algorithm does not alter the assignment of ORs and CPs to biclusters, but merely provides better visualization of the biclusters.

6.5 References

1. Firestein S: **How the olfactory system makes sense of scents**. *Nature* 2001, **413**:211-218.
2. Mombaerts P: **Genes and ligands for odorant, vomeronasal and taste receptors.** *Nat Rev Neurosci* 2004, **5**:263–278.
3. Glusman G, Yanai I, Rubin I, Lancet D: **The complete human olfactory subgenome**. *Genome Res* 2001, **11**:685–702.
4. Olender T, Fuchs T, Linhart C, Shamir R, Adams M, Kalush F, Khen M, Lancet D: **The Canine Olfactory Subgenome**. *Genomics* 2004, **83**:361-372.
5. Zhang X, Firestein S: **The olfactory receptor gene superfamily of the mouse.** *Nat Neurosci* 2002, **5**:124-133.
6. Niimura Y, Nei M: **Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages.** *Gene* 2005, **346**:23-28.
7. Niimura Y, Nei M: **Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods**. *Proc Natl Acad Sci U S A* 2005, **102**:6039–6044.
8. Aloni R, Olender T, Lancet D: **Ancient genomic architecture for mammalian olfactory receptor clusters.** *Genome Biol* 2006, **7**:R88.
9. Liu AH, Zhang X, Stolovitzky GA, Califano A, Firestein SJ: **Motif-based construction of a functional map for mammalian olfactory receptors**. *Genomics* 2003, **81**:443–456.
10. Solan Z, Horn D, Ruppin E, Edelman S: **Unsupervised learning of natural languages**. *in Proc Natl Acad Sci* 2005, **102**:11629-11634.
11. Kunik V, Solan Z, Edelman S, Ruppin E, Horn D: **Motif Extraction and Protein Classification**. *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)* 2005.
12. Segal L, Lapidot M, Solan Z, Ruppin E, Pilpel Y, Horn D: **Nucleotide variation of regulatory motifs may lead to distinct expression patterns.** *Bioinformatics* 2007, **23**(13):i440-i449.
13. **AmphibiaWeb: Information on amphibian biology and conservation. [web application]. Berkeley, California, http://amphibiaweb.org/.** 2008.
14. Tinsley RC, Kobel HR: **.** In: *The Biology of Xenopus.* Oxford: Oxford Scientific Press.; 1996: pp 41-43.
15. Ciccarelli FD, Doerks T, Mering Cv, Creevey CJ, Snel B, Bork P: **Toward Automatic Reconstruction of a Highly Resolved Tree of Life**. *Science* 2006, **311**(5765):1283 - 1287.
16. Maestrini E, Monaco AP, McGrath JA, Ishida-Yamamoto A, Camisa C, Hovnanian A, Weeks DE, Lathrop M, Uitto J, Christiano AM: **A molecular defect in loricrin, the major component of the cornified cell envelope, underlies Vohwinkel's syndrome**. *Nature Genetics* 1996, **13**:70-77.
17. Rouquier S, Blancher A, Giorgi D: **The olfactory receptor gene repertoire in primates and mouse: Evidence for reduction of the functional fraction in primates**. *PNAS* 2000, **97**(6):2870–2874.
18. Lagerström MC, Hellström AR, Gloriam DE, Larsson TP, Schiöth HB, Fredriksson R: **The G Protein–Coupled Receptor Subset of the Chicken Genome**. *PLoS Comput Biol* 2006, **2**(6):e54.
19. Bjarnadóttir TK, Gloriam DE, Hellstrand SH, Kristiansson H, Fredriksson R, Schiöth HB: **Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse**. *Genomics* 2006, **88**:263-273.
20. Parker M, Wong Y, Parker S: **An ion-responsive motif in the second transmembrane segment of rhodopsin-like receptors.** *Amino Acids* 2008, **Online**.
21. Menashe I, Aloni R, Lancet D: **A probabilistic classifier for olfactory receptor pseudogenes**. *BMC Bioinformatics* 2006, **7**:393.
22. Alioto TS, Ngai J: **The odorant receptor repertoire of teleost fish**. *BMC Genomics* 2005, **6**:173.

23. Freitag J, Ludwig G, Andreini I, Rössler P, Breer H: **Olfactory receptors in aquatic and terrestrial vertebrates**. *J Comp Physiol* 1998, **183**(5):635-650.
24. Meroz Y, Horn D: **Biological roles of specific peptides in enzymes.** *Proteins: Structure, Function, and Bioinformatics* 2008, **Online**.
25. Warren WC, Hillier LW, Graves JAM, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT *et al*: **Genome analysis of the platypus reveals unique signatures of evolution**. *Nature* 2008, **453**(7192):175-183.
26. Sheu JJ-C, Hua C-H, Wan L, Lin Y-J, Lai M-T, Tseng H-C, Jinawath N, Tsai M-H, Chang N-W, Lin C-F *et al*: **Functional Genomic Analysis Identified Epidermal Growth Factor Receptor Activation as the Most Common Genetic Event in Oral Squamous Cell Carcinoma**. *Cancer Res* 2009, **69**:2568.
27. Tso C-L, Shintaku P, Chen J, Liu Q, Liu J, Chen Z, Yoshimoto K, Mischel PS, Cloughesy TF, Liau LM *et al*: **Primary Glioblastomas Express Mesenchymal Stem-Like Properties**. *Mol Cancer Res* 2006, **4**:607.
28. Sasaki H, Yu C-Y, MeiruDai, Tam C, Loda M, Auclair D, Chen LB, Elias A: **Elevated serum periostin levels in patients with bone metastases from breast but not lung cancer**. *Breast Cancer Res and Treatment* 2003, **77**:245–252.
29. Arnason U, dagger JA, Adegoke D, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X *et al*: **Mammalian mitogenomic relationships and the root of the eutherian tree.** *Proceedings of the National Academy of Science* 2002, **99**:8151-8156.
30. Lunter G: **Dog as an Outgroup to Human and Mouse**. *PLoS Comput Biol* 2007, **3**(4):e74.
31. Günther HS, Schmidt NO, Phillips HS, Kemming D, Kharbanda S, Soriano R, Modrusan Z, Meissner H, Westphal M, Lamszus K: **Glioblastoma-derived stem cell-enriched cultures form distinct subgroups according to molecular and phenotypic criteria**. *Oncogene* 2008(2897–2909).
32. Man O, Gilad Y, Lancet D: **Prediction of the odorant binding site of olfactory receptor proteins by human–mouse comparisons**. *Protein Sci* 2004, **13**:240-254.
33. Dhillon IS: **Co-clustering documents and words using bipartite spectral graph partitioning**. *In Proceedings of the ACM SIGKDD Conference* 2001:269-274.

6.6 Supplementary Material

Supplementary Tables and figures are also found in http://adios.tau.ac.il/ORPS/ and in http://www.biomedcentral.com/1471-2148/9/91/additional/

6.6.1 GPCR remote homologies

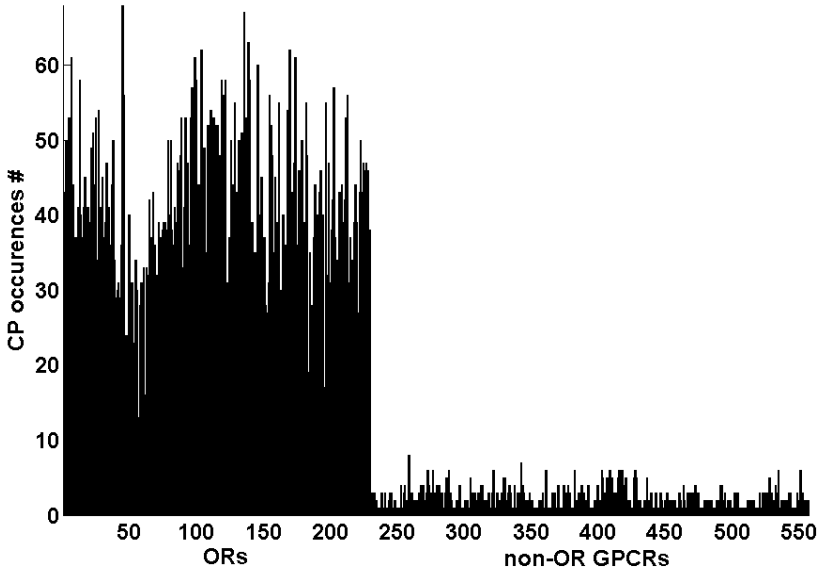


Figure S1. The number of CP occurrences (hits) for each of the chicken 229 intact and pseudogene ORs and 281 non-OR GPCR from [1]

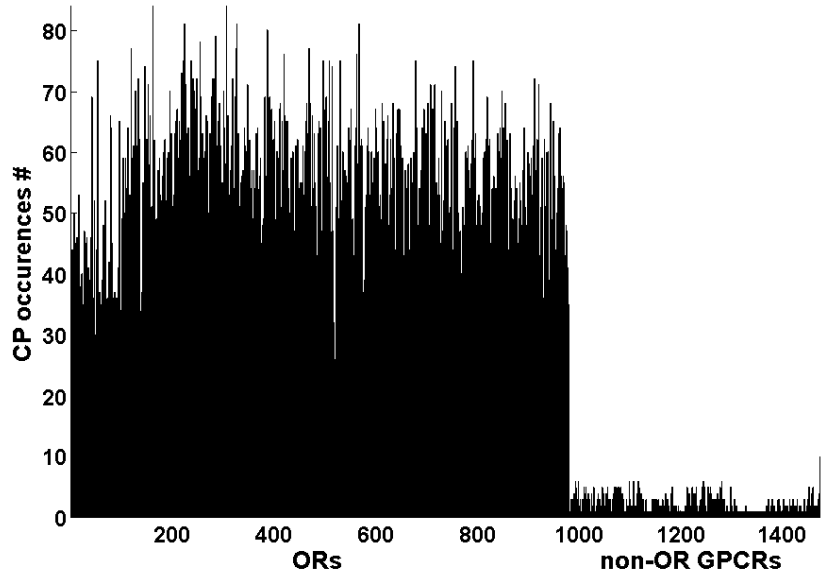


Figure S2. The number of CP occurrences (hits) for each of the mouse 978 intact ORs and 386 non-OR GPCR from [2].

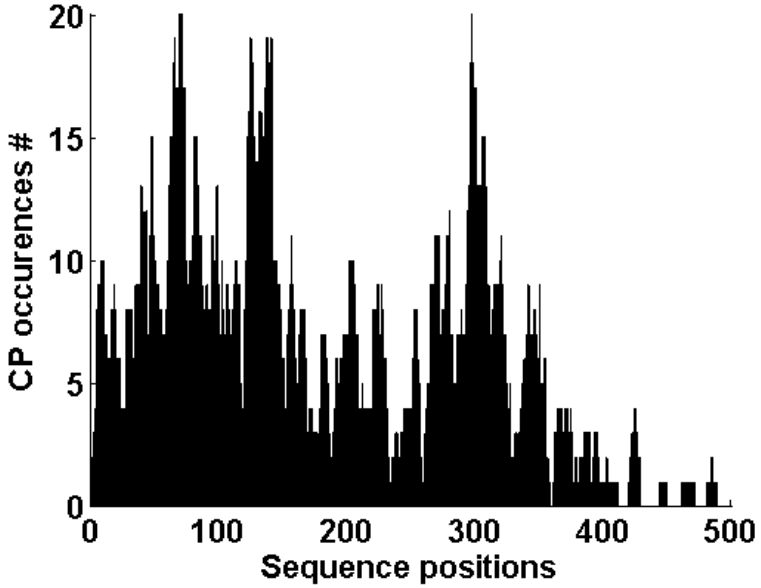


Figure S3. CP coverage of amino acid positions along chicken non-OR GPCR sequences. The positions are shown up to 500 amino-acids for clarity.

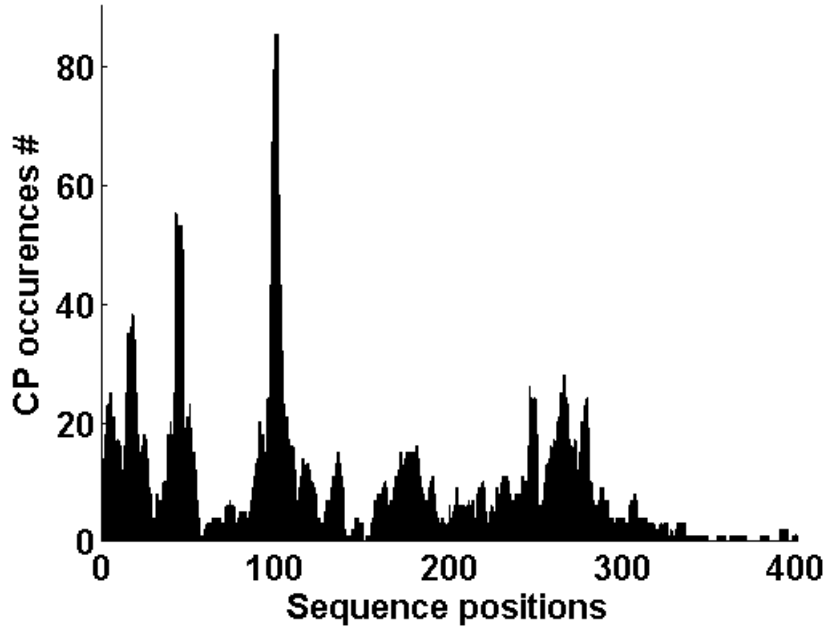


Figure S4. CP coverage of amino acid positions along mouse non-OR GPCR sequences. The positions are shown up to 400 amino-acids for clarity.

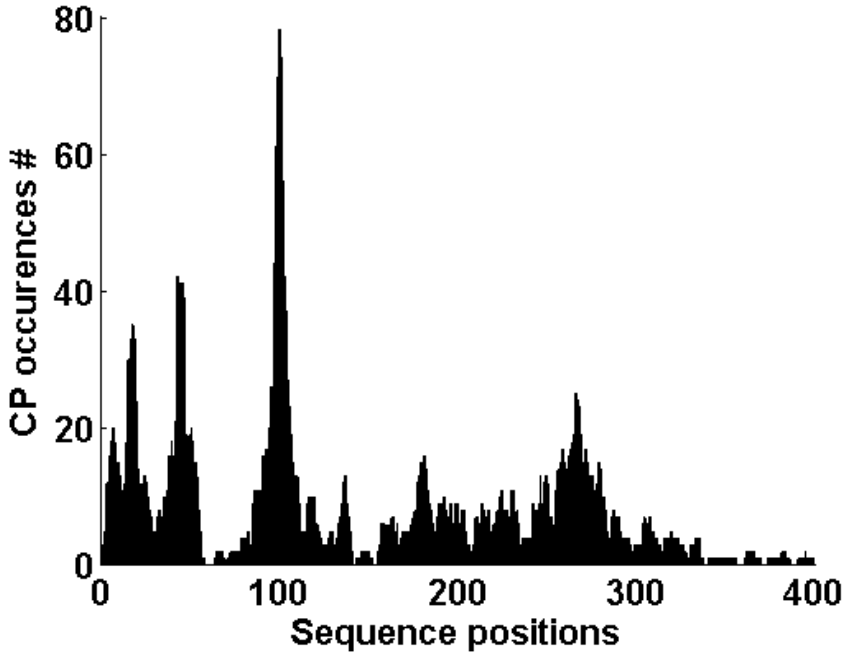


Figure S5. CP coverage of amino acid positions along mouse non-OR GPCR sequences. The positions are shown up to 400 amino-acids for clarity.

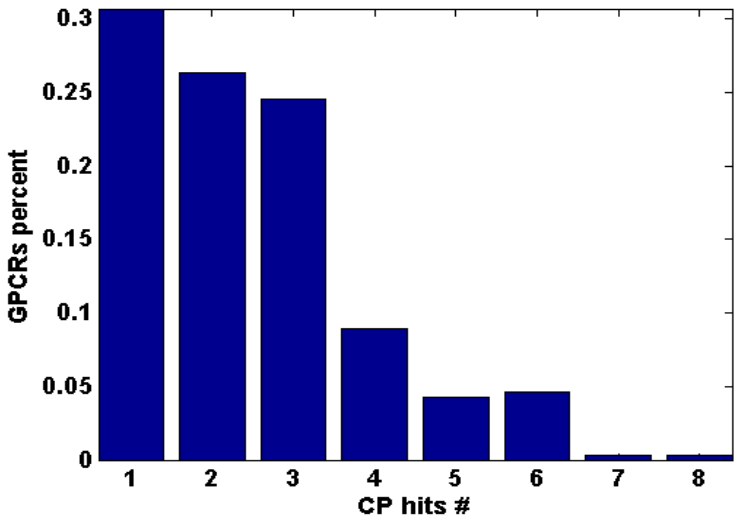


Figure S6. Histogram of the percent of chicken non-OR GPCRs as a function of the number of CPs occurring in them.

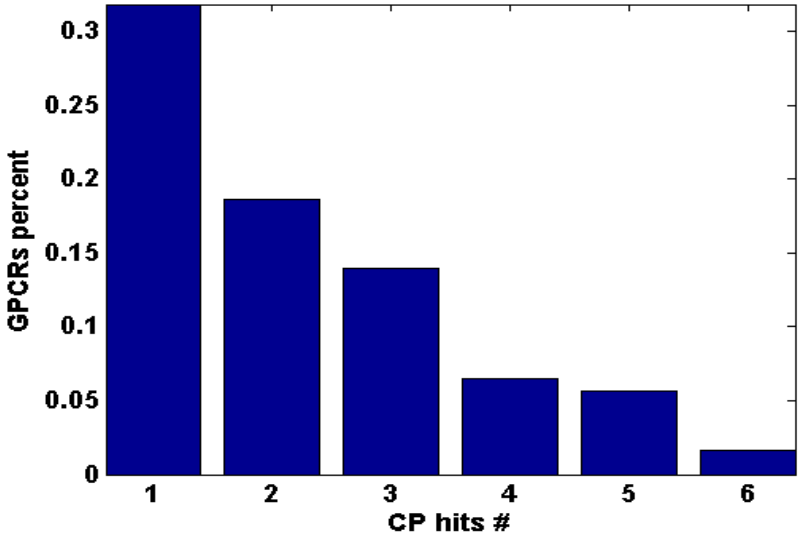


Figure S7. Histogram of the percent of mouse non-OR GPCRs as a function of the number of CPs occurring in them.

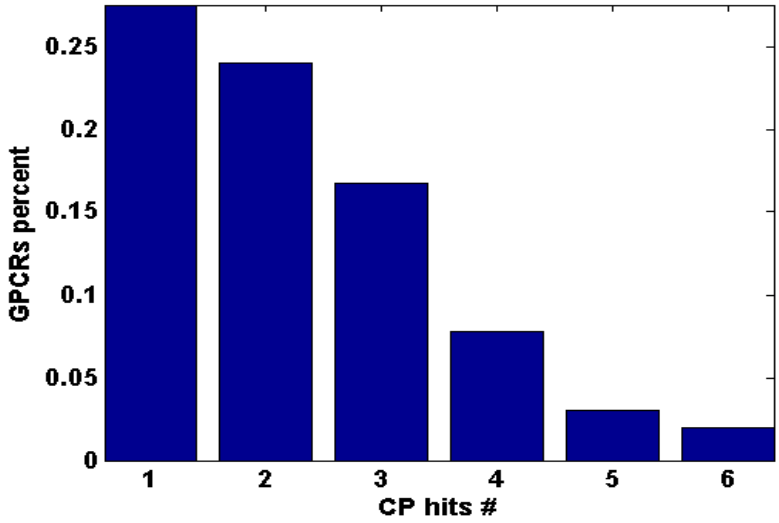


Figure S8. Histogram of the percent of human non-OR GPCRs as a function of the number of CPs occurring in them.

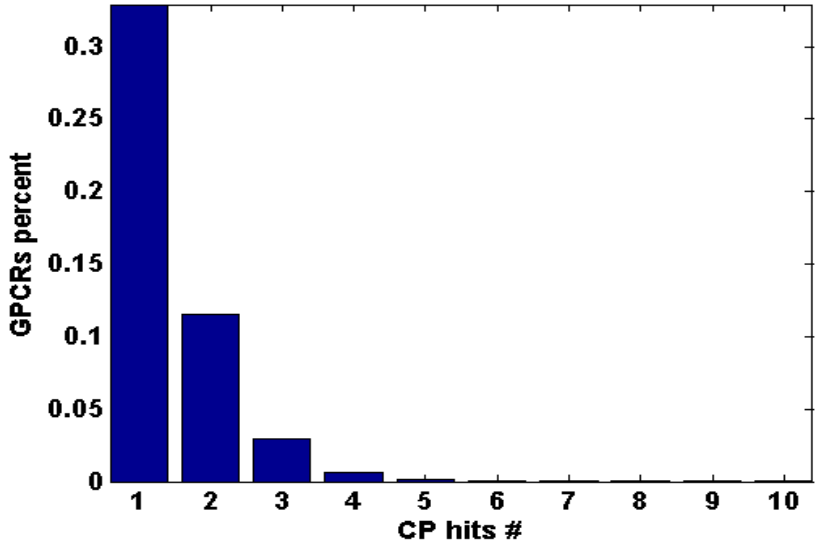


Figure S9. Histogram of the percent of human and mouse randomly permuted non-OR GPCRs as a function of the number of CPs occurring in them.

6.6.2 Locations of CPs on the OR sequence

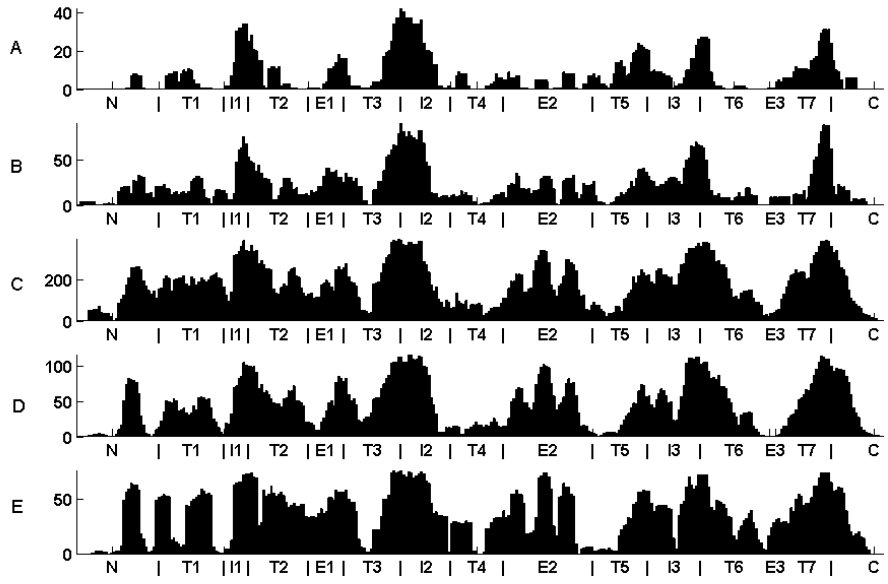


Figure S10. Coverage of ORs by CPs as a function of positions along the OR sequence for Pufferfish (A), Zebrafish (B), Frog (C), Lizard (D) and Chicken (E). Positions start from the N-terminal (N), through Transmembrane domains 1-7 (T1-T7), Intracellular loops (I1-I3) and extracellular loops (E1-E3), ending with the C-terminal (C).

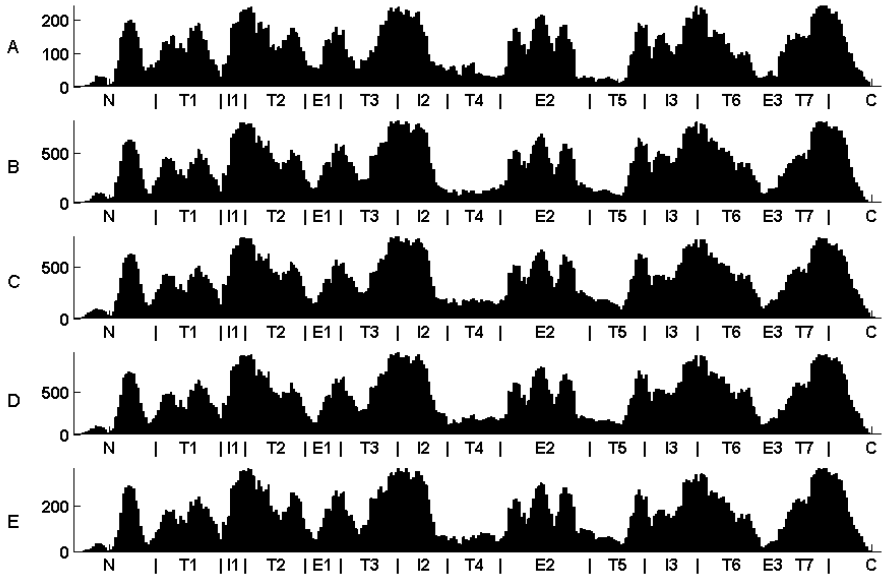


Figure S11. Coverage of ORs by CPs as a function of positions along the OR sequence for Platypus (A), Opossum (B), Dog (C), Mouse (D) and Human (E). Positions are ordered using the same coordinates as in Figure S10.

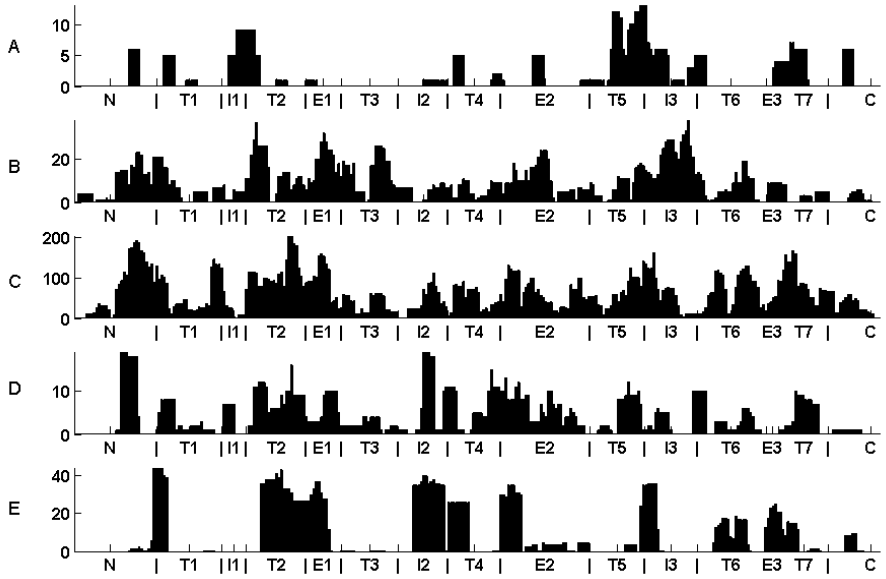


Figure S12. Coverage of ORs by CPs as a function of positions along the OR sequence for *novel* CPs of Pufferfish (A), Zebrafish (B), Frog (C), Lizard (D) and Chicken (E). Positions are ordered using the same coordinates as in Figure S10.

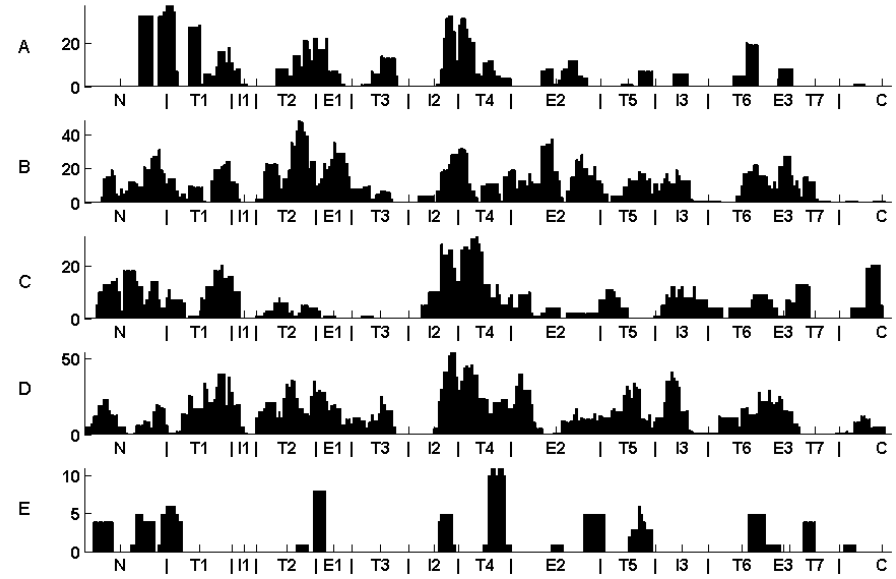


Figure S13. Coverage of ORs by CPs as a function of positions along the OR sequence for *novel* CPs of Platypus (A), Opossum (B), Dog (C), Mouse (D) and Human (E). Positions are ordered using the same coordinates as in Figure S10

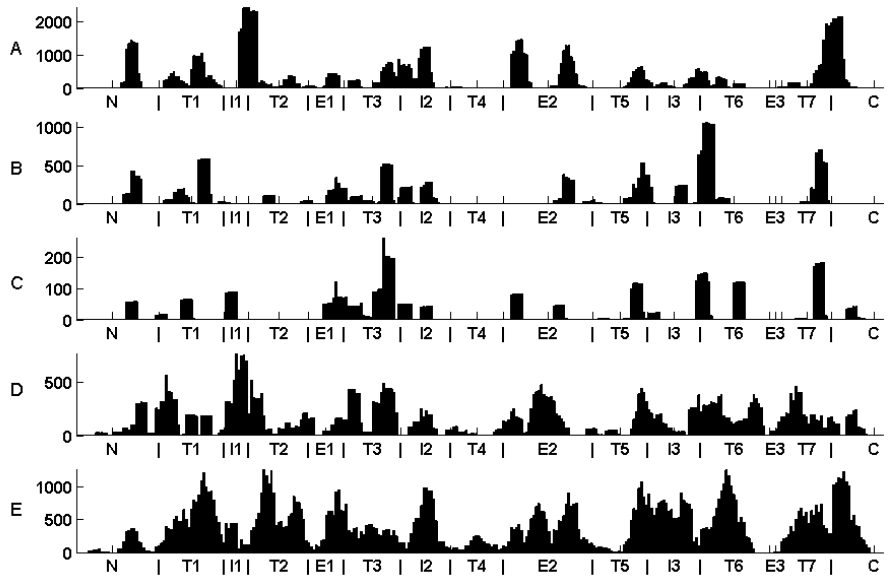


Figure S14. Coverage of ORs by CPs as a function of positions along the OR sequence for CPs *lost* in Pufferfish (A), Zebrafish (B), Frog (C), Lizard (D) and Chicken (E). Positions are calculated over all ORs other than the specific species and ordered using the same coordinates as in Figure S10.

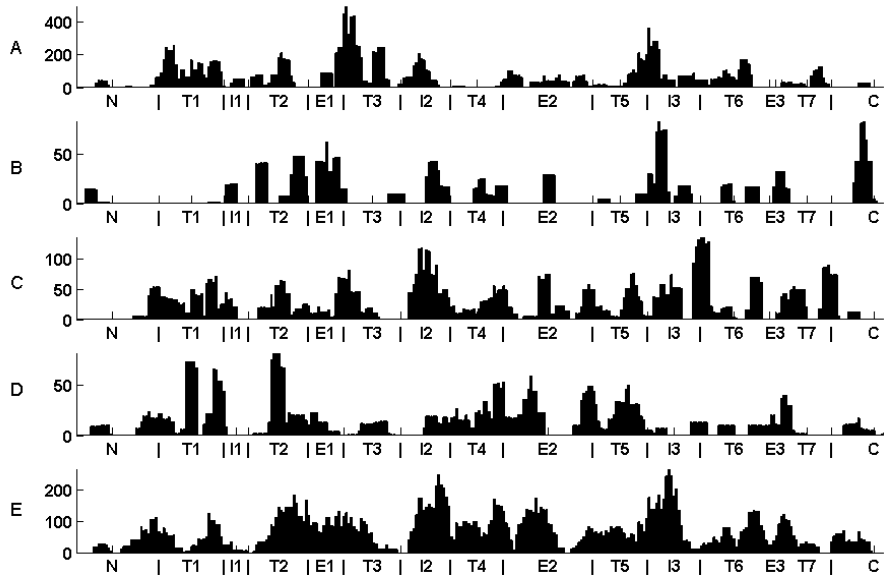


Figure S15. Coverage of ORs by CPs as a function of positions along the OR sequence for CPs *lost* in Platypus (A), Opossum (B), Dog (C), Mouse (D) and Human (E). Positions are calculated over all ORs other than the specific species and ordered using the same coordinates as in Figure S10

6.6.3 CP-space reveals internal clusters

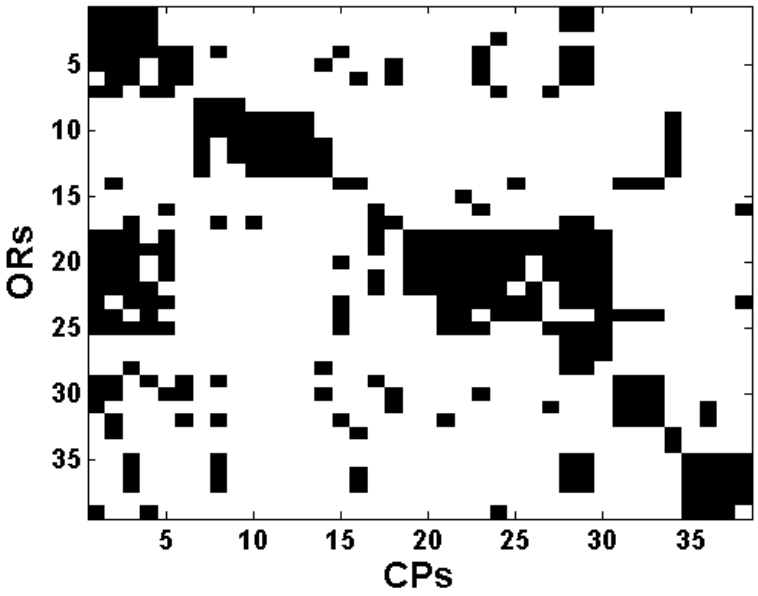


Figure S16. Biclustering results of Pufferfish. Y-axis corresponds to chicken ORs and X-axis to CPs novel to pufferfish.

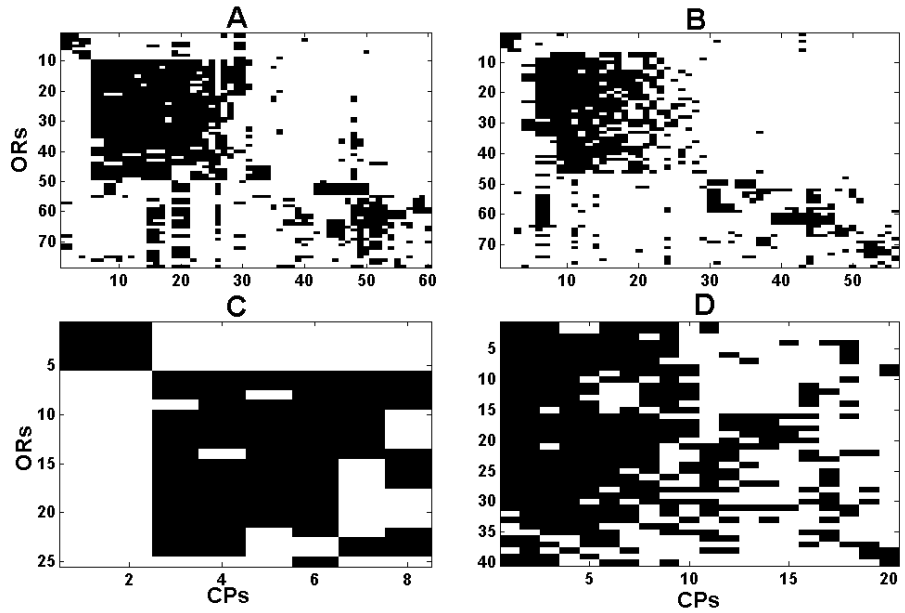


Figure S17. Biclustering results of Chicken. Y-axis corresponds to chicken ORs and X-axis to CPs novel to the MRCA of fish and tetrapods (A), tetrapods ancestor (B), amniotes ancestor (C) and CPs novel to chicken (D).

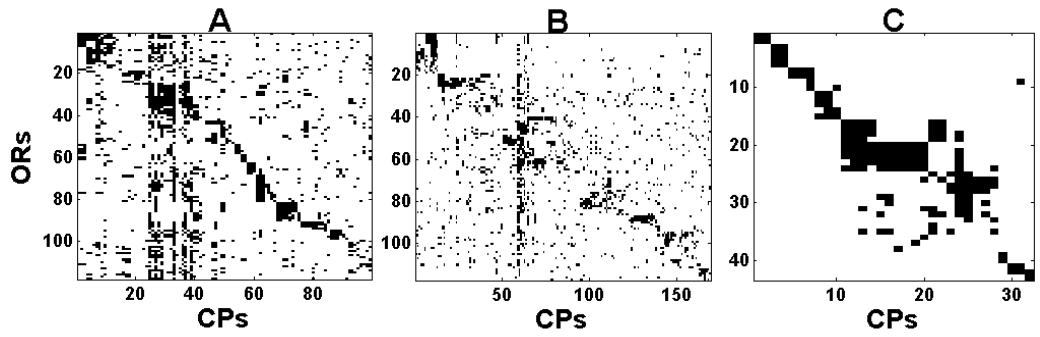


Figure S18. Biclustering results of Lizard. Y-axis corresponds to lizard ORs and X-axis to CPs novel to the MRCA of fish and tetrapods (A), ancestor of tetrapods (B) and CPs novel to lizard (C).

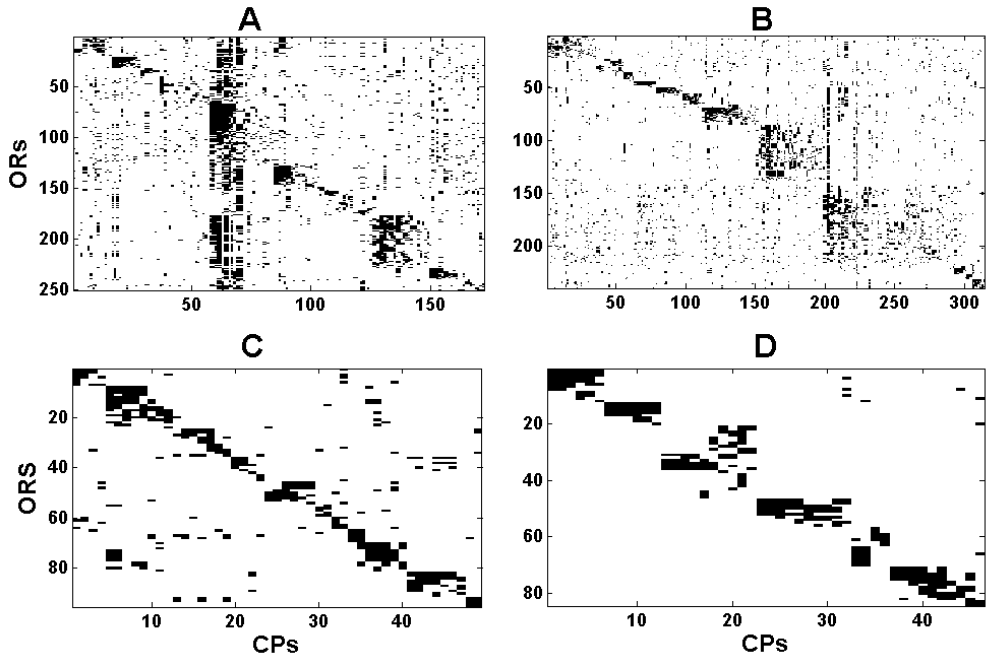


Figure S19. Biclustering results of Platypus. Y-axis corresponds to platypus ORs and X-axis to CPs novel to MRCA of fish and tetrapods (A), ancestor of tetrapods (B), ancestor of amniotes (C) and ancestor of mammals (D) CPs.

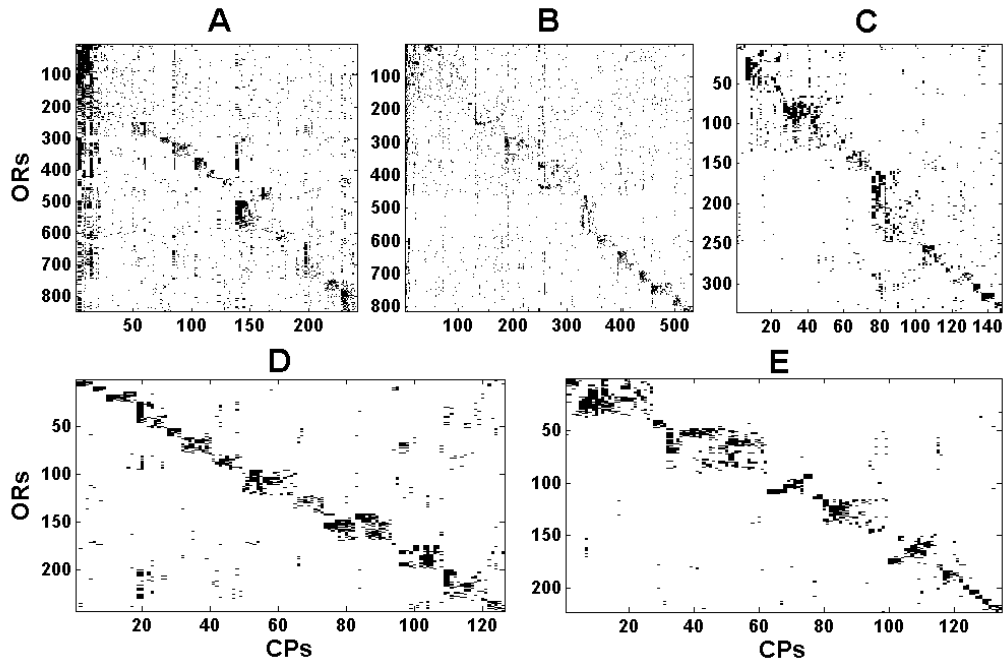


Figure S20. Biclustering results of Opossum. Y-axis corresponds to opossum ORs and X-axis to CPs novel to MRCA of fish and tetrapods (A), ancestor of tetrapods (B), ancestor of amniotes (C), ancestor of mammals (D) and ancestor of marsupials (E).

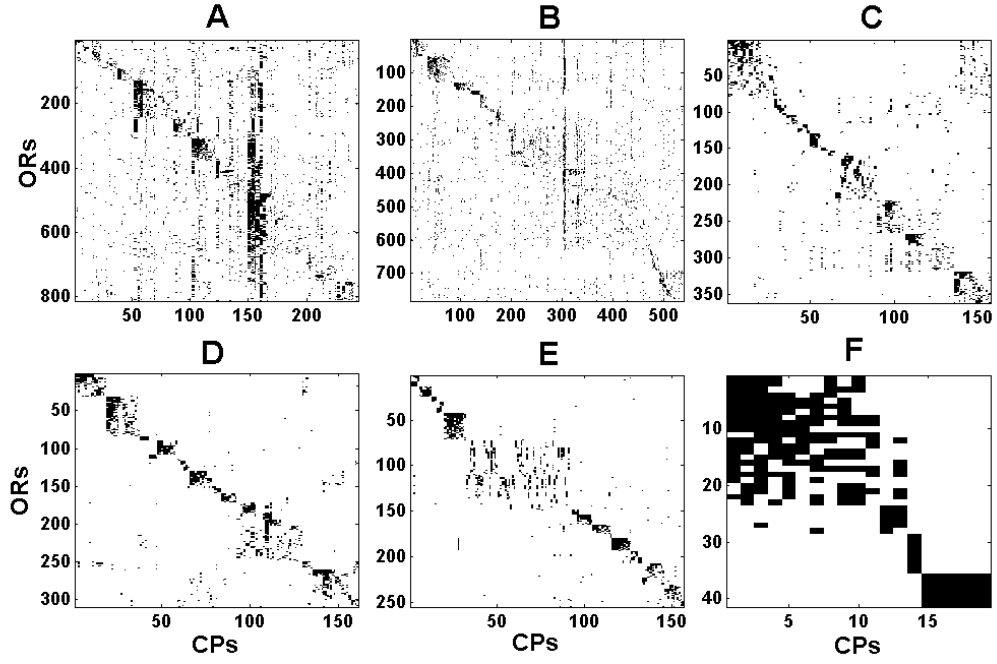


Figure S21. Biclustering results of Dog. Y-axis corresponds to dog ORs and X-axis to CPs novel to MRCA of fish and tetrapods (A), ancestor of tetrapods (B), ancestor of amniotes (C), ancestor of mammals (D), ancestor of marsupials (E) and ancestor of eutherians (F).

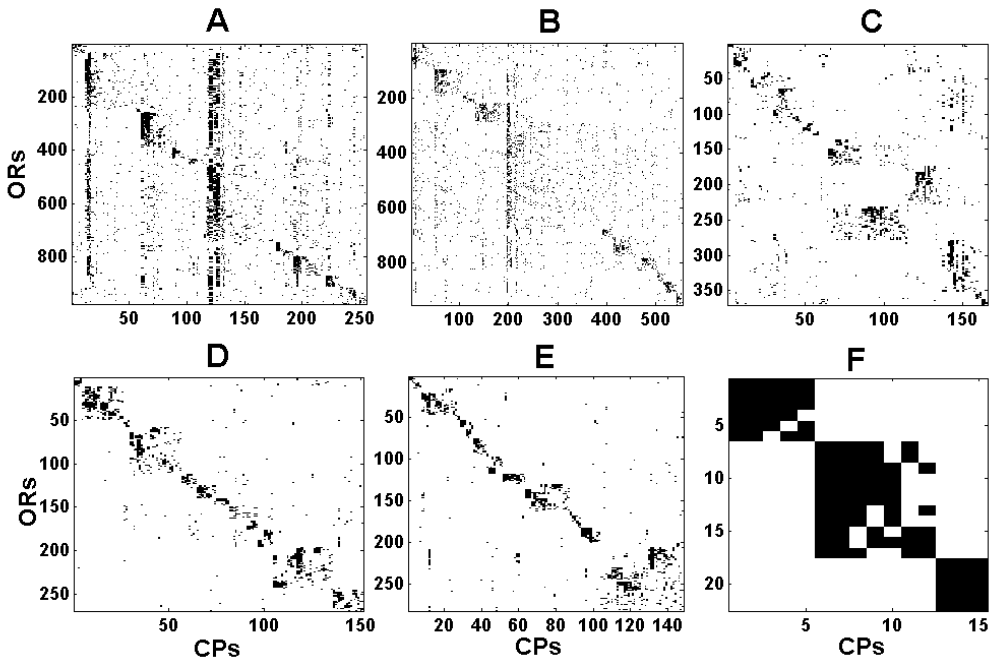


Figure S22. Biclustering results of Mouse Y-axis corresponds to mouse ORs and X-axis to CPs novel to MRCA of fish and tetrapods (A), ancestor of tetrapods (B), ancestor of amniotes (C), ancestor of mammals (D), ancestor of marsupials (E) and CPs novel to mouse (F).

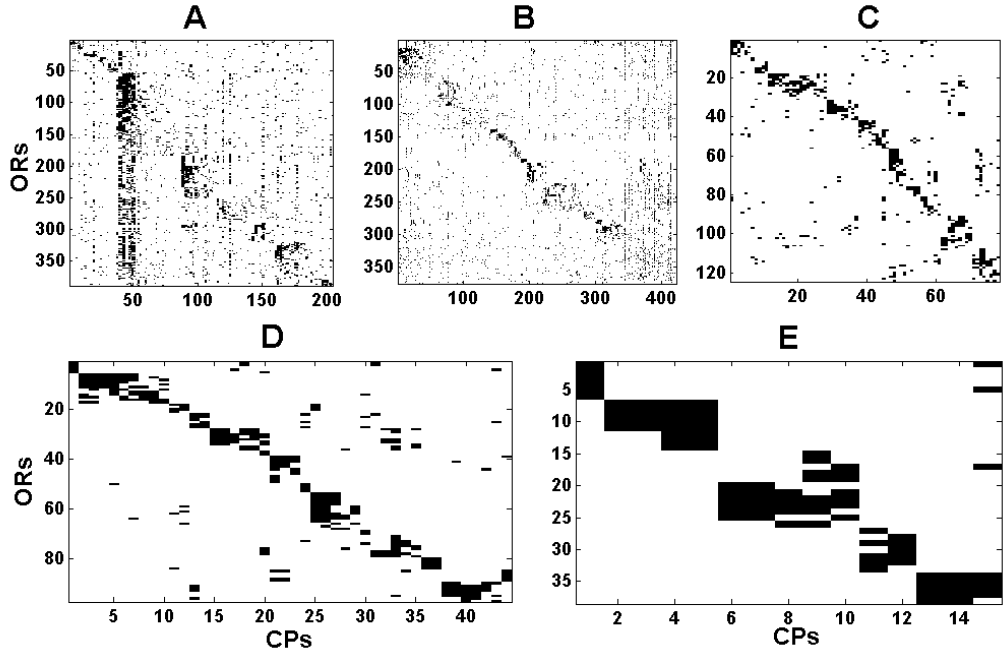


Figure S23. Biclustering results of Human. Y-axis corresponds to human ORs and X-axis to CPs novel to MRCA of fish and tetrapods (A), ancestor of tetrapods (B), ancestor of amniotes (C), ancestor of mammals (D) and ancestor of marsupials (E).

Chapter 7

Analysis of aminoacyl tRNA synthetases using Common Peptides[9]

7.1 Introduction

The aminoacyl-tRNA synthetases (aaRSs) are key participants in the translation mechanism of the cell, catalyzing the esterification of specific amino acids and their corresponding tRNAs. They have drawn attention in recent years due to their crucial function. Extensive study [1, 2, 3 , 4] has been done on their structure in order to understand the exact mechanism by which they operate. Their key role in the heart of the translation process and their connection to the genetic code make them natural candidates for evolutionary studies, aiming to pinpoint the way translation has started in the hypothesized primitive cell and the way it evolved to the current stage [5, 6]. The aaRS fall into two classes based on the topology of their ATP binding domains.

We study the aaRS families by using the Common Peptides (CPs) methodology, which has been successfully employed for olfactory receptors [7]. CPs are extracted by applying the Motif Extraction algorithm (MEX) [8, 9] to each of the aaRS families. Their lists are then combined to provide a unified CP list, which forms our system of reference. Representing aaRS sequences in this CP space, we compare different aaRS families and track evolutionary relations between them. We put special emphasis on uncovering relationships between mitochondria and the three kingdoms of life. We find novel class-determining signature CPs, possibly bearing functional roles. We indicate the most ancient CPs, based on the reconstruction of the CPs on the tree of life (ToL) and show that abundant CPs have functional importance by showing that most of them occupy known catalytic and binding sites on PDB, while some others have undetermined functionality.

In essence, we provide a novel perspective, regarding aaRS families through the use of CPs, and point out novel CPs that may reside on functional locations.

7.2 Methods

7.2.1 Data

We analyze 5406 sequences belonging to 22 different enzyme families of aminoacyl tRNA synthetases corresponding to Enzyme Commision (EC) number 6.1.1.x from Enzyme and UniProt

[9] Based on the paper *Analysis of aminoacyl tRNA synthetases using Common Peptides,* Assaf Gottlieb, Milana Frenkel-Morgenstern, Mark Safro, David Horn, in preparation.

databases. Table 1 lists the different aaRSs and number of sequences in each. 6.1.1.x family, including synthetases of the 20 common amino acids (AA), along with the uncommon pyrrolysine and O-phosphoseryl-tRNA ligase (SepRS).

biotin-[acetyl-CoA carboxylase] synthetase (birA) sequences were also studied because of their similarity to aaRSs. 1664 birA sequences were downloaded from UniProt.

7.2.2 Method of Common Peptides

The data downloaded from Enzyme database contains some almost identical sequences, belonging either to very close species or different strains of the same species. In order not to identify Common Peptides (CPs) that are common only in the sense that they exist mainly on these near-identical sequences, we used single linkage clustering with a threshold of 90% sequence identity to filter these groups, keeping only their central representatives (defined in terms of average closeness to other cluster members). The remaining sequences thus represent a 'non-redundant set' of the Enzyme database within EC 6.1.1.

We followed a procedure similar to [7]. This procedure starts by applying the unsupervised Motif EXtraction algorithm (MEX) [8], [10] to each of the 22 non-redundant sets of enzyme sequences, thus leading to.22 separate sets of Common Peptides (CPs), of length 5 amino-acids or more. The separate lists of CPs are then unified, removing redundancy from the unified list by removing CPs containing smaller CPs. The unified list contains 10612 CPs. Finally, all 10612 CPs are searched on all aaRS sequences (including aaRSs where the CP was not extracted by MEX). We thus end up with a CP space in which all the sequences are represented.

7.2.3 Assignment of proteins to kingdoms

Assignment of species to the different kingdoms of life (archaea, bacteria and eukarya), including separation of mitochondrial sequences into a separate group was done automatically using Kyoto Encyclopedia of Genes and Genomes (KEGG) Organisms [11-13], Karyn's Genomes [14] and Ciccarelli tree of life [15] followed by manual curation.

Table 24. Properties of aaRS, ordered by class. Displayed are number of sequences used in the MEX analysis, number of CPs derived from each aaRS category, the total number of CP hits from the unified list of 10612 CPs and the number of CPs found only within a given family.

EC	Name	Class	# of sequences	# of MEX CPs	# of observed CPs	# of specific CPs

6.1.1.1	Tyrosyl tRNA synthetase	I	261	400	758	239
6.1.1.2	Tryptophanyl tRNA synthetase	I	121	163	323	102
6.1.1.4	Leucyl tRNA synthetase	I	344	1031	1730	591
6.1.1.5	Isoleucyl tRNA synthetase	I	271	871	1608	568
6.1.1.9	Valyl tRNA synthetase	I	211	641	1293	378
6.1.1.10	Methionyl tRNA synthetase	I	248	634	1121	386
6.1.1.16	Cysteinyl tRNA synthetase	I	362	505	998	301
6.1.1.17	Glutaminyl tRNA synthetase	I	373	645	1237	407
6.1.1.18	Glutamyl tRNA synthetase	I	37	96	178	50
6.1.1.19	Arginyl tRNA synthetase	I	327	677	1275	421
6.1.1.3	Threonyl tRNA synthetase	II	279	671	1128	431
6.1.1.6	Lysyl tRNA synthetase	II	192	340	651	175
6.1.1.7	Alanyl tRNA synthetase	II	193	506	1019	314
6.1.1.11	Seryl tRNA synthetase	II	345	489	874	264
6.1.1.12	Aspartatyl tRNA synthetase	II	294	586	992	341
6.1.1.14	Glycyl tRNA synthetase	II	226	432	773	265
6.1.1.15	Prolyl tRNA synthetase	II	369	792	1313	475
6.1.1.20	Phenylalanyl tRNA synthetase	II	495	682	1576	418
6.1.1.21	Histidyl tRNA synthetase	II	312	393	838	246
6.1.1.22	Asparaginyl tRNA synthetase	II	124	213	402	130
6.1.1.n2	O-Phosphoseryl-tRNA synthetase	II	17	31	87	20
6.1.1.26	Pyrrolysyl tRNA synthetase	II	5	16	20	11

7.2.4 Fitting CPs to the tree of life and phylogenetic analysis

We used the tree of life (ToL) constructed by [15] to follow relationships between species. Being interested in the upper nodes of the tree, species were mapped to the ToL also by genus name when the specific species was not found in [15].This left us with 2293 sequences that could be mapped. The assessment of CP origins uses the Wagner parsimony, as implemented by the Phylogeny Inference Package computer programs PHYLIP.

7.3 Results

7.3.1 Frequent CPs

CPs that occur on multiple sequences reside on conserved regions, which are naturally assumed to play structural or functional role. Table 2 displays the top 10 occurring CPs. All of them occur exclusively in the more conserved class I aminoacyl-tRNA synthetases. The first CP in Table 2, KMSKS, is one of two well-known signature sequence motifs related to class I aminoacyl-tRNA synthetases that catalyze the amino acid activation with ATP [16, 17]. We note that the second well-known signature, HIGH, is not included because our MEX application was limited to motifs of length 5 or more. Some CPs, however, contain this motif or a mutated form of it; the most abundant, residing in the top 30 frequent CPs are LHMGH and HIGHA, occurring on 249 (4.6%) and 242 (4.5%) sequences respectively.

It is interesting to note that most of the frequent CPs do not occur at all (or occur in negligible amounts) in Eukaryotes, although existing in their mitochondria. Further discussion of the differences in CP representation in different kingdoms is found in the section "Evolutionary Aspects of CPs".

Only one full sequence does not have any CP hits (5 other sequence fragments ranging between 10-49 AA long also do not have hits). This is a type-2 seryl-tRNA synthetase of *M. thermophila (strain DSM 6194 / PT)*, a rare form unique to methanogen archaea [5]. The other 6 methanogen archaea species containing this rare form have 2-7 CP hits on them.

Table 25. Top 10 most frequent CPs. Each row displays the number of sequences the CP occurs in (percent of all sequences), # of aaRSs it appears in, and the number of occurrences in each kingdom +mitochondria (percent of all the sequences belonging to this kingdom)

CP	sequence occurrences (percentage)	# of aaRSs	bacteria (percentage)	eukarya (percentage)	archaea (percentage)	mitochondria (percentage)
KMSKS	1364 (25%)	9	1196 (26.4%)	24 (14.5%)	102 (17.1%)	42 (39.6%)
KSLGN	502 (9%)	9	440 (9.7%)	5 (3.0%)	40 (6.7%)	17 (16.0%)
ISRQR	345 (6%)	3	296 (6.5%)	0 (0.0%)	36 (6.0%)	13 (12.3%)
GRPGWH	333 (6%)	1	297 (6.5%)	6 (3.6%)	24 (4.0%)	6 (5.7%)
PSPTG	329 (6%)	2	319 (7.0%)	3 (1.8%)	0 (0.0%)	7 (6.6%)
FPHHE	327 (6%)	1	294 (6.5%)	0 (0.0%)	28 (4.7%)	5 (4.7%)
PYANG	318 (6%)	2	295 (6.5%)	0 (0.0%)	18 (3.0%)	5 (4.7%)
RQRYWG	310 (6%)	2	283 (6.2%)	0 (0.0%)	22 (3.7%)	5 (4.7%)

SKSKG	299 (6%)	8	266 (5.9%)	1 (0.6%)	32 (5.4%)	0 (0.0%)
PYPSG	294 (5%)	2	284 (6.3%)	0 (0.0%)	4 (0.7%)	6 (5.7%)

7.3.2 CPs as Class Signatures

CPs are generally not specific to a particular aaRS, but some appear dominantly in class I or class II synthetases (see Table 1 for classification of the aaRSs).

The number of CPs in each class is summarized in the Venn diagram in figure 1.

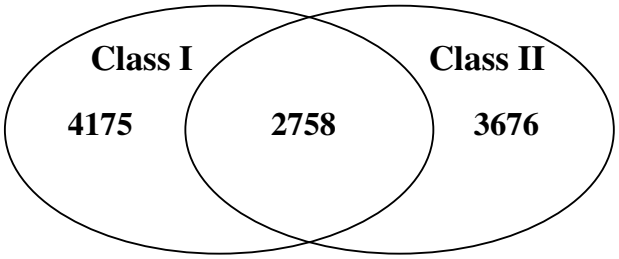


Figure 1. Relative abundance of CPs in class I and class II synthetases. 11 CPs are specific to pyrrolysine tRNAPyl synthetase and thus do not belong to any class.

Table 3 lists CPs that display preference for one of the classes and cover more than half of the different aaRSs. Shown are the number of different enzymes that a specific CP appears on, and the number of different aaRSs. Variations of the known class I signatures (HIGH and KMSKS) were omitted.

These CPs may be used to aid classification task (see [9]). Furthermore, these CPs may signify a functional or structural constrained region, related to the specific type of operation of each class enzymes. While class I has two known signatures, class II have none. In this respect, the CPs specific to class II appearing in Table 3, may be regarded as novel signatures.

Table 26.: Novel CP class signatures for Aminoacyl-tRNA synthetases. GVERL may be part of more general class II motifs [3, 18].

CP	# of class I aaRSs	# of class I occurrences	# of class II aaRSs	# of class II occurrences
TADEI	8	47	1	1
ALADE	8	37	1	2
KSLGN	7	500	2	2

SKSKG	7	296	1	3
SKGNV	7	178	1	4
DVIAR	7	73	0	0
DVVAR	7	60	1	2
ADAIR	7	38	1	1
GLDLL	7	35	1	1
GVERL	0	0	8	92
DLVEE	1	2	7	67
GLDRI	1	1	7	43
AEAVL	1	2	7	24
ERISA	0	0	7	24
LRLAE	0	0	6	38
AAGVR	2	2	6	47

7.3.3 CPs as Features

CPs span a space in which the aaRS sequences are represented. In this space, we calculate Pearson correlations between different aaRSs. A heat map of these correlations is presented in figure 2, where the aaRSs are grouped according to their classes.

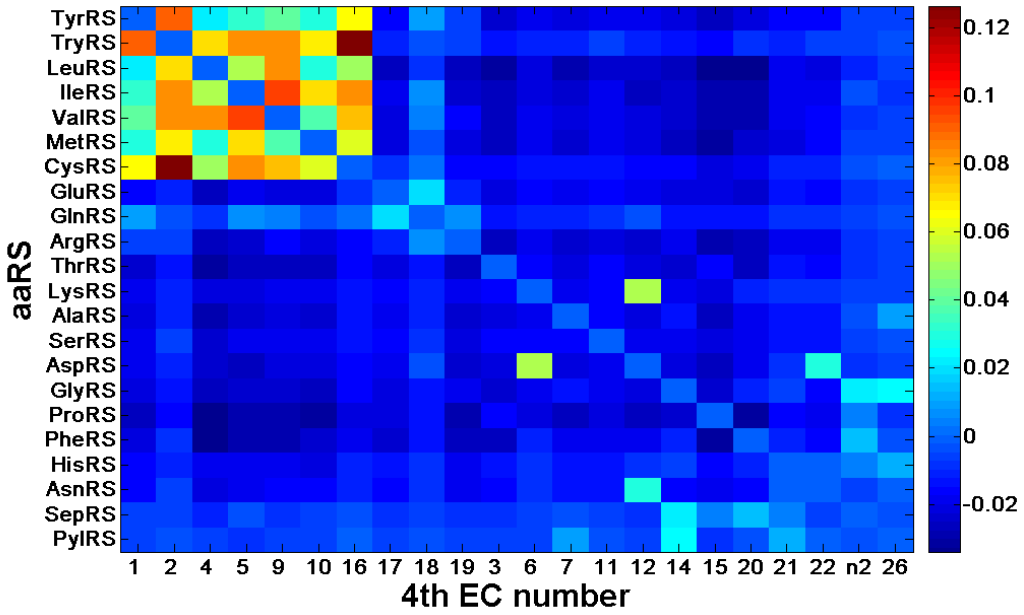


Figure 2. Heat map of Pearson cross-correlations of different aaRSs according to their shared CPs. Self correlations were left out for the purpose of clearer presentation.

While the absolute values of the correlations are small, some correlations stand out above the background. Class I aaRSs are generally very close to one another, except for GluRS, GlnRS and ArgRS. In class II correlations are much smaller, but for those between LysRS and AspRS and, to some extent, AspRS and AsnRS. Interestingly, SepRS and PylRS show above background correlation to GlyRS.

Interesting relations emerge when the correlations are calculated for each kingdom separately (mitochondria taken as a separate kingdom). Bacteria are dominant and the correlations between aaRSs calculated on them show wide resemblance to figure 2.

aaRS correlations calculated among eukaryotes show only three aaRS pairs that stand out above the background . The first pair is GluRs and GlnRS (see [19] for related discussion). The second pair is ProRS and GluRS and the third observed pair contains ProRS and GlnRS that belong to different classes. These pairs of high-correlation do not exhibit such behavior in bacteria. Furthermore, the correlation values are much higher than observed in other kingdoms (0.31 for GluRs and GlnRS and 0.49 for ProRS and GluRS). The high correlation of eukaryotic ProRS and GluRS matches the observation made by [20], pointing out that the genes of ProRS and GluRS are organized differently in the three kingdoms of the tree of life. In bacteria and archaea, distinct genes encode the two proteins while in several organisms from the eukaryotic phylum of coelomate metazoans, the two polypeptides are carried by a single polypeptide chain to form a bifunctional protein, postulated to result from a gene fusion event.

Correlations between aaRSs within archaea show class I correlations that are similar to bacteria but for two exceptions; TryRS only correlates with TyrRS, and MetRS has no correlations with other class I members (MetRS is indeed mentioned as having lower level of similarity in [21]).

Last, calculating correlations between aaRSs using only the mitochondria exhibits a slightly different pattern, in which class I aaRSs do not correlate with TyrRS and class II PheRS correlates with AlaRS, a correlation that does not appear when calculated in other kingdoms. The correlation heat-maps for each kingdom are found in the supplementary figures S1-S4.

7.3.4 Evolutionary Aspects of CPs

It has been demonstrated by [7] that reconstructing CPs onto a phylogenetic tree can track interesting evolutionary events. Following the same philosophy, we first examine the assignment to the different kingdoms of life, separating mitochondria from Eukaryotes. Table 4 displays the relative abundance of CPs in each kingdom.

Table 27. Statistics of each kingdom of life, separating mitochondria from the eukaryotes.

Kingdom	# of proteins	Percentage of CPs	Specific CPs out of all observed in kingdom
bacteria	4538	94.7%	62.8%
eukarya	166	13.1%	3.7%
archaea	596	28.9%	13.5%
Mitochondria	106	11.1%	1.3%

According to Ciccarelli's Tree of Life (ToL) [15], its first few branches define various archaea genuses, and only then the tree splits into bacteria and eukarya. Accordingly, we define distinct sets of CPs appearing in all three kingdoms, in the joint node of bacteria and Eukaryotes and in each of the 3 kingdoms exclusively, and analyze the distribution of each such set among the different aaRSs. This is displayed in Figure 3, enabling us to study the history of aaRS formation.

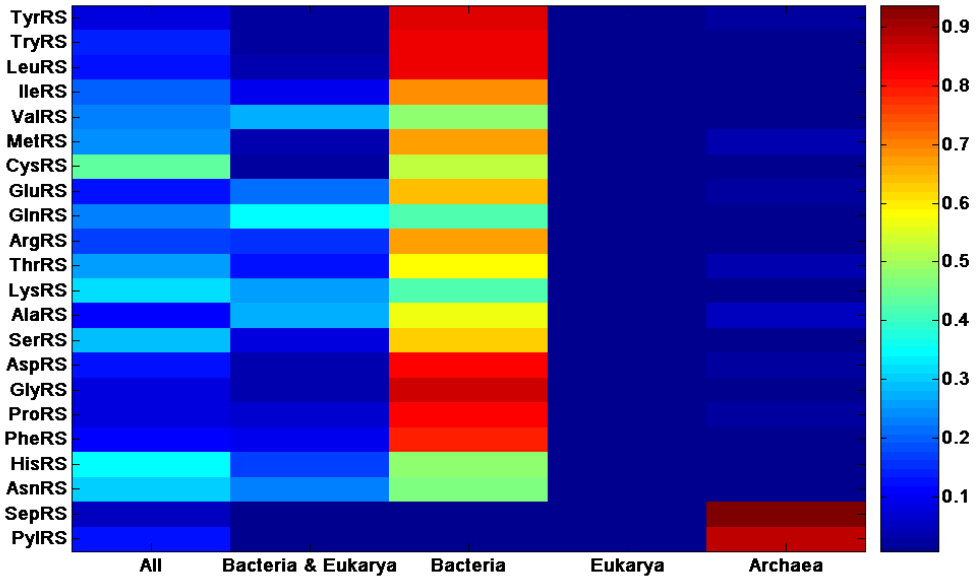


Figure 3. Distribution of different aaRS families according to the CP groups appearing in all 3 kingdoms, only in bacteria and eukarya, and in each kingdom exclusively.

Figure 3 shows that for the 20 common amino acids, most of the CPs originate only in bacteria, while for SepRS and PylRS, most originate in archaea. CysRS is an exception, having many of its CPs belonging to all 3 kingdoms. This is consistent with [5] who pointed out that CysRS provides considerable evidence of interdomain horizontal gene transfer, particularly involving archaea. Other interesting observations emerge, easily viewed when bacteria-specific CPs are excluded from figure 3 (see supplementary figure S5). It becomes then obvious that ValRS, AlaRS, GluRs and GlnRs, are

relatively more conserved in bacteria and eukarya than in all 3 kingdoms. These observations are in accordance with the observations made by [5].

Using Ciccarelli's Tree of Life (ToL) [15], we reconstruct the CPs on the tree using parsimony (see methods). Of particular interest are the CPs that occur highest in the ToL, i.e. in branches that include archaea. 15 CPs that occur highest in the tree (i.e. nodes including archaeal species) and also appear in more than half of the species in the tree are listed in table 6. All of them belong predominantly to class I. 6 of them are also in the top ranked CPs list (table 2). These CPs appear highest in the tree, hence they are highly conserved. Being also exclusive to one of the classes (with few exceptions) suggests that they are good candidates for functional regions and should be subjected to further exploration. Four CPS from this list (CGGTH, EVETP, GGRYD and the known class I signature KMSKS) also appear on sequences having resolved structures in PDB, and will be discussed in the next section. Only two out of the four overlap known functional regions (GGRYD overlaps atp-binding region and KMSKS ligand binding region).

Table 28. CPs occurring in the top nodes of the ToL (including archaeal sequences) and covering more than half of the ToL species. Bold-faced CPs are specific instantiations of class I signature motifs.

CP	# of class I aaRSs	# of class I occurrences	# of class II aaRSs	# of class II occurrences
DWCISRQ	3	256	0	0
DVLDTW	1	165	0	0
GRPGWH	1	333	0	0
TTTPWT	1	255	0	0
CGGTH	1	1	2	140
EVETP	0	0	5	205
FPHHE	1	327	0	0
GGRYD	0	0	1	253
ISRQR	3	345	0	0
KMSKS	8	1362	1	2
KSLGN	7	500	2	2
LHIGH	10	216	1	2
SKSKG	7	296	1	3
TTRPE	3	192	0	0
WTTTP	1	259	0	0

7.3.5 Mitochondria

Mitochondria constitute a special case, residing in Eukaryotic cells, yet bearing similarity to bacteria [22]. The Venn diagram in figure 4 displays the percentage of CPs that appear in mitochondrial sequences according to the way they are shared by aaRS sequences from other kingdoms. Furthermore, the heat map in figure 4 shows the correlation of each of the kingdoms to each other, with mitochondria and bacteria showing a high similarity. The closeness to bacteria, as postulated by [23] is clearly observed.

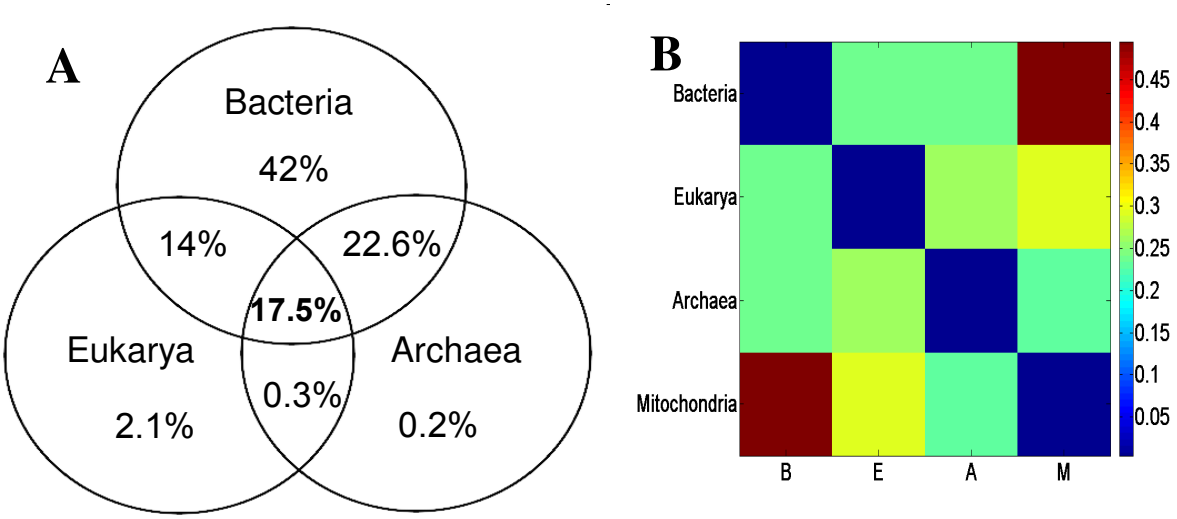


Figure 4. (A) Venn diagram of the percentage of CPs present in mitochondrial sequences shared with other kingdoms. (B) Heat map of correlation between different kingdoms + mitochondria. Self correlations were left out for the purpose of clearer presentation.

15 CPs are found to be specific to Mitochondria. They are listed in table 5. It is interesting to note that the CP "QQQQQ" that appears only in mitochondria, appears in isoleucyl, leucyl and histidil tRNA synthetases. This CP usually appears more than once in a sequence and typically it is part of a longer stretch of glutamines. This may point out that these proteins contain intrinsically unstructured regions (IURs) [24 , 25].

Table 29. CPs specific to the Mitochondria

CP	# of enzymes the CP appears in
TTPIFYVN	9

SLESGH	7
VHSHW	7
ELADALGGLLNRCTA	5
QWGNYFLH	5
STWELLD	5
KIQQAA	5
CVRQTNGFVQRHAPWKL	4
ITNCGSGF	4
YKALEAVS	4
GTLLQPV	4
KLPEFNR	4
AVQHFW	4
QQQQQ	4
VLQWL	4

7.3.6 Biological role

We looked for occurrences of our CPs on resolved structures within the PDB database. We restricted ourselves to CPs that cover more than half of the sequences of at least one aaRS family, henceforth termed *frequent CP*s. A large number of the frequent CPs occupies binding sites that can be read-off these structures. Their list is presented in Table 7. 78% of the frequent CPs appearing in table 7 overlap known catalytic and binding regions.

Table 30. Occurrence of frequent CPs in aaRSs that on sequences with a PDB entry. Most of them occur at strategic sites, of high relevance to the biological function. KMSKS is a known class I motif.

CP	Uniprot ID	PDB ID	remarks
AADIL	SYW_BACST	1I6K	ligand binding domain
CGGTH	SYA_AQUAE	1YFR	-
DDTNP	SYQ_DEIRA	1EUQ	tRNA binding area
DFQAR	SYS_AQUAE	2DQ3	ligand-binding motif
DPRMPT	SYQ_DEIRA	1EUQ	-
DTGMG	SYA_AQUAE	1YFR	ATP binding domain
EISSCS	SYS_AQUAE	2DQ3	ligand-binding motif
EVETP	SYK1_ECOLI	1BBU	-
FEMLGN	SYA_AQUAE	1YFR	ATP binding domain
FIGKD	SYM_PYRAB	1F4L	ligand-binding motif
FQARR	SYS_AQUAE	2DQ3	ligand-binding motif

FRNEG	SYK1_ECOLI	1BBU	262R-catalytic residue, ligand binding region
GDYFK	SYA_AQUAE	1YFR	ligand-binding region
GEIIGGS	SYN_PYRHO	1X54	ATP and metal binding region
GFGLG	SYN_PYRHO	1X54	ligand-binding motif
GGGRY	SYH_THEAC	1KMN	ATP-binding region
GGRYD	SYH_THEAC	1KMN	ATP-binding region
GGSQRE	SYN_PYRHO	1X54	ATP and metal binding region
GIGIDR	SYK1_ECOLI	1BBU	480R-catalytic residue, ligand binding region
GMGLE	SYA_AQUAE	1YFR	ATP-binding domain
GPCGP	SYA_AQUAE	1YFR	metal binding domain
GRGYV	SYA_AQUAE	1YFR	-
GVIHW	SYQ_DEIRA	1EUQ	tRNA binding area
HHTFF	SYA_AQUAE	1YFR	ATP-binding domain
HNPEF	SYK1_ECOLI	1BBU	ligand binding region
KAFYM	SYN_PYRHO	1X54	ligand-binding motif
KLSKR	SYE1_THEMA	2O5R	metal-binding region
KMSKS	SYW_BACST	1I6K	motif of class I: ligand binding domain
LDLRR	SYD_SULTO	1WYD	-
LNGSG	SYS_AQUAE	2DQ3	ligand-binding motif
LRAKI	SYQ_DEIRA	1EUQ	tRNA binding area
LRFDF	SYA_AQUAE	1YFR	-
LRIEDT	SYE1_THEMA	2O5R	-
MGCYG	SYP_ENTFA	2J3L	ligand-binding motif
NGSGLA	SYS_AQUAE	2DQ3	ligand-binding motif
PPHGG	SYD_PYRKO	1B8A	ligand-binding region
PSPTG	SYE1_THEMA	2O5R	ATP-binding region
PTAEV	SYS_AQUAE	2DQ3	ligand-binding motif
PTHEE	SYP_ENTFA	2J3L	ligand-binding motif
PYANG	SYM_PYRAB	1F4L	ligand-binding motif
QLPKF	SYS_AQUAE	2DQ3	-
REISS	SYS_AQUAE	2DQ3	ligand-binding motif
RFAPSP	SYE1_THEMA	2O5R	ATP-binding region
RIEDTD	SYE1_THEMA	2O5R	-
SFGDY	SYA_AQUAE	1YFR	ligand-binding region
SKRKL	SYQ_DEIRA	1EUQ	catalytic region
TAEVP	SYS_AQUAE	2DQ3	ligand-binding motif
TLNGS	SYS_AQUAE	2DQ3	ligand-binding motif
TRFPP	SYQ_DEIRA	1EUQ	catalytic region
TYGFP	SYA_AQUAE	1YFR	-
VHTLN	SYS_AQUAE	2DQ3	ligand-binding motif

WDDPR	SYQ_DEIRA	1EUQ	-
YDRLF	SYQ_DEIRA	1EUQ	-

7.3.7 biotin-[acetyl-CoA carboxylase] synthetase (birA) and aaRSs

Biotin-[acetyl-CoA carboxylase] synthetase (birA) is a bifunctional protein, acting as biotin-protein synthetase and binding to DNA to regulate its own transcription. [26] demonstrates structural similarity between its active sites and class II aaRS, although no sequence similarity exists. It is thus of interest to find out whether certain CPs are common to birA, revealing local similarities that may be related to the structural similarity in their binding sites.

In order to select dominant CPs, mutual to aaRSs and birA, we first performed the same procedure described in the method of common peptides section for birA sequences, i.e. we extracted a new set of 1630 non-redundant birA CPs. We next chose only CPs that appear in both aaRS and birA lists, either by exact match or by inclusion, where we chose the included CP (i.e. birA CPs could be part of a larger aaRS CP and vice-versa). This has lead to a list of 28 CPs that appear in both lists (either exactly or being part of a larger CP in one of the two lists). By requiring appearance in a minimum of 20 sequences in both aaRSs and birA, we filtered out four CPs, listed in table 8. The most prominent CP is GILIE (appearing in birA also as GILVE or GILTE). It covers more than a hundred sequences from both aaRSs and birA, appears dominantly in class II aaRSs. According to PDB it resides on a catalytic site in birA and appears in close vicinity (2 residues apart) of a ligand binding AC2 region in ThrRS (PDB IDs 1bia [27] for birA and 1qf6 [28] for ThrRS).

Table 31. frequent CPs common to aaRSs and birA. Alternatives in brackets marks one-mutation far CPs that where selected on birA but not in aaRSs.

CP (alternatives)	Structural properties	# of aaRSs occurrences	# of birA occurrences	# of class I occurrence	# of class II occurrences

				s	
GILIE (GILVE, GILTE)	Catalytic site in birA, AC2 with lignad residue in ThrRS	2	198	2	113
GALRL (GALLL)	α-helix in AspRS	1	32	0	42
GEALG (GETLG)	Helix-turn-Helix in birA	4	22	1	31
LRAAL	α-helix in birA	13	86	43	7

7.4 Discussion

In this paper, we employ the Common Peptides (CPs) methodology to analyze aminoacyl tRNA synthetases (aaRSs).

The CPs allow us to discover novel class I and class II aaRSs signatures, allowing for further research examining the role of these signatures in the function of the two different aaRS classes.

Using the CPs as feature space in which aaRSs are expressed, we are able to identify correlations between the aaRSs. These correlations, calculated for sequences belonging to species from a single kingdom (e.g. bacteria, archaea, eukaryotes and mitochondria) reveal differences in the aaRS correlations between different kingdoms.

Using [15] tree of life (ToL), we are able to allocate the CPs to different branches using parsimony. This reveals which CPs have older origins and thus are the most conserved ones across kingdoms, suggesting a functional or structural role for these "ancient" CPs. Focusing on mitochondria, we are able to show that mitochondria and bacteria are undoubtedly much closer to each other than mitochondria to other kingdoms. We also identify mitochondria-specific CPs.

Next we assess the biological significance of frequently occurring CPs by checking whether they overlap known binding and catalytic regions for sequences having a PDB structure. We show that the majority (80%) of the frequently occurring CPs overlap such regions (p_value>0.023, corresponding to FDR>=0.05)

Last we find CPs common to both aaRSs and biotin-[acetyl-CoA carboxylase] synthetase (birA) which have a structural catalytic region resemblance; although no sequence similarity is present. We identify four such CPs that are candidates for analysis that could verify whether they constitute a sequential region that allows for the structural similarity.

In essence, using CPs to analyze aaRSs provides a novel point of view on aaRSs relations, evolution and similarity to other proteins.

7.5 References

1. Perona JJ, Rould MA, Steitz TA: **Structural basis for transfer RNA aminoacylation by Escherichia coli glutaminyl-tRNA synthetase**. *Biochemistry* 1993, **32**(34):8758-8771.
2. Delarue M: **Aminoacyl-tRNA synthetases**. *Curr Opin Struct Biol* 1995, **5**(1):48-55.
3. Burbaum JJ, Schimmel P: **Structural relationships and the classification of aminoacyl-tRNA synthetases**. *J Biol Chem* 1991, **266**(26):16965-16968.
4. Arnez JG, Moras D: **Structural and functional considerations of the aminoacylation reaction**. *Trends Biochem Sci* 1997, **22**(6):211-216.
5. Woese CR, Olsen GJ, Ibba M, Soll D: **Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process**. *Microbiol Mol Biol Rev* 2000, **64**(1):202-236.
6. Delarue M: **Partition of aminoacyl-tRNA synthetases in two different structural classes dating back to early metabolism: implications for the origin of the genetic code and the nature of protein sequences**. *J Mol Evol* 1995, **41**(6):703-711.
7. Gottlieb A, Olender T, Lancet D, Horn D: **Common peptides shed light on evolution of Olfactory Receptors**. *BMC Evol Biol* 2009, **9**:91.
8. Solan Z, Horn D, Ruppin E, Edelman S: **Unsupervised learning of natural languages**. *in Proc Natl Acad Sci* 2005, **102**:11629-11634.
9. Kunik V, Solan Z, Edelman S, Ruppin E, Horn D: **Motif Extraction and Protein Classification**. *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05)* 2005.
10. Kunik V, Meroz Y, Solan Z, Sandbank B, Weingat U, Ruppin E, Horn D: **Functional representation of enzymes by specific peptides**. *PLOS Comp Biol* 2007, **3**(8):e167.
11. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
12. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Res* 2006, **34**(Database issue):D354-357.
13. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T *et al*: **KEGG for linking genomes to life and the environment**. *Nucleic Acids Res* 2008, **36**(Database issue):D480-484.
14. **Karyn's Genomes - http://www.ebi.ac.uk/2can/genomes/genomes.html**.
15. Ciccarelli FD, Doerks T, Mering Cv, Creevey CJ, Snel B, Bork P: **Toward Automatic Reconstruction of a Highly Resolved Tree of Life**. *Science* 2006, **311**(5765):1283 - 1287.
16. Austin J, First EA: **Comparison of the catalytic roles played by the KMSKS motif in the human and Bacillus stearothermophilus trosyl-tRNA synthetases**. *J Biol Chem* 2002, **277**(32):28394-28399.
17. Kobayashi T, Takimura T, Sekine R, Kelly VP, Kamata K, Sakamoto K, Nishimura S, Yokoyama S: **Structural snapshots of the KMSKS loop rearrangement for amino acid activation by bacterial tyrosyl-tRNA synthetase**. *J Mol Biol* 2005, **346**(1):105-117.
18. Eriani G, Delarue M, Poch O, Gangloff J, Moras D: **Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs**. *Nature* 1990, **347**(6289):203-206.
19. Siatecka M, Rozek M, Barciszewski J, Mirande M: **Modular evolution of the Glx-tRNA synthetase family--rooting of the evolutionary tree between the bacteria and archaea/eukarya branches**. *Eur J Biochem* 1998, **256**(1):80-87.
20. Berthonneau E, Mirande M: **A gene fusion event in the evolution of aminoacyl-tRNA synthetases**. *FEBS Lett* 2000, **470**(3):300-304.
21. Nagel GM, Doolittle RF: **Phylogenetic analysis of the aminoacyl-tRNA synthetases**. *J Mol Evol* 1995, **40**(5):487-498.
22. Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR: **Mitochondrial origins**. *Proc Natl Acad Sci U S A* 1985, **82**(13):4443-4447.
23. Margulis L: **Symbiosis in Cell Evolution**: W.H.Freeman & Co Ltd; 1981.
24. Simon M, Hancock JM: **Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins**. *Genome Biol* 2009, **10**(6):R59.
25. Tompa P: **Intrinsically unstructured proteins evolve by repeat expansion**. *Bioessays* 2003, **25**(9):847-855.
26. Artymiuk PJ, Rice DW, Poirrette AR, Willet P: **A tale of two synthetases**. *Nat Struct Biol* 1994, **1**(11):758-760.
27. Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW: **Escherichia coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains**. *Proc Natl Acad Sci U S A* 1992, **89**(19):9257-9261.
28. Sankaranarayanan R, Dock-Bregeon AC, Romby P, Caillet J, Springer M, Rees B, Ehresmann C, Ehresmann B, Moras D: **The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site**. *Cell* 1999, **97**(3):371-381.

7.6 Supplementary Material

Supplementary Tables and figures are also found in http://adios.tau.ac.il/aaRSCP/

7.6.1 CPs as features

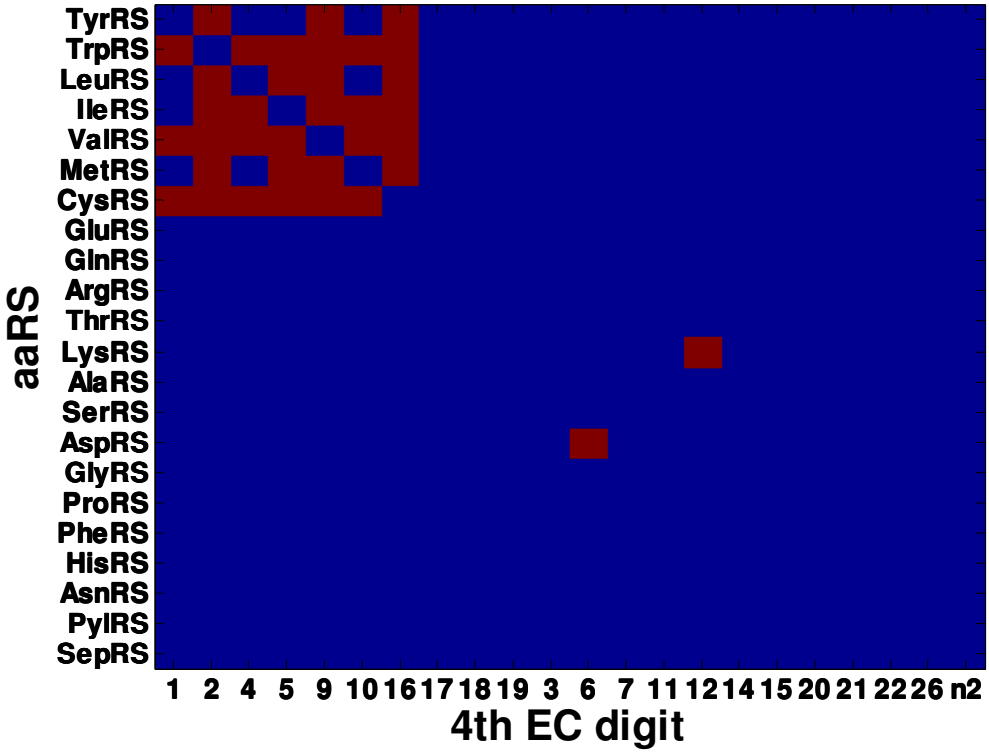


Figure S1. Pearson cross-correlations of different aaRSs according to their shared CPs. Only correlations with p-value <0.01 are shown in red. Self correlations were left out for the purpose of clearer presentation.

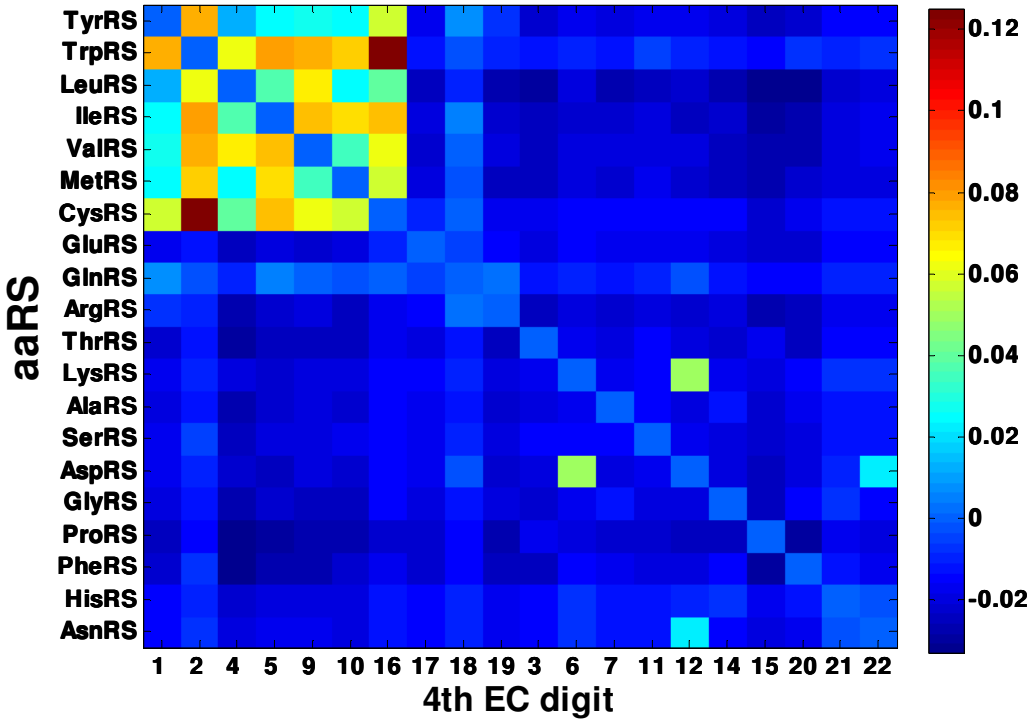


Figure S2. Heat map of Pearson cross-correlations of different aaRSs according to their shared CPs in Bacteria. Self correlations were left out for the purpose of clearer presentation.

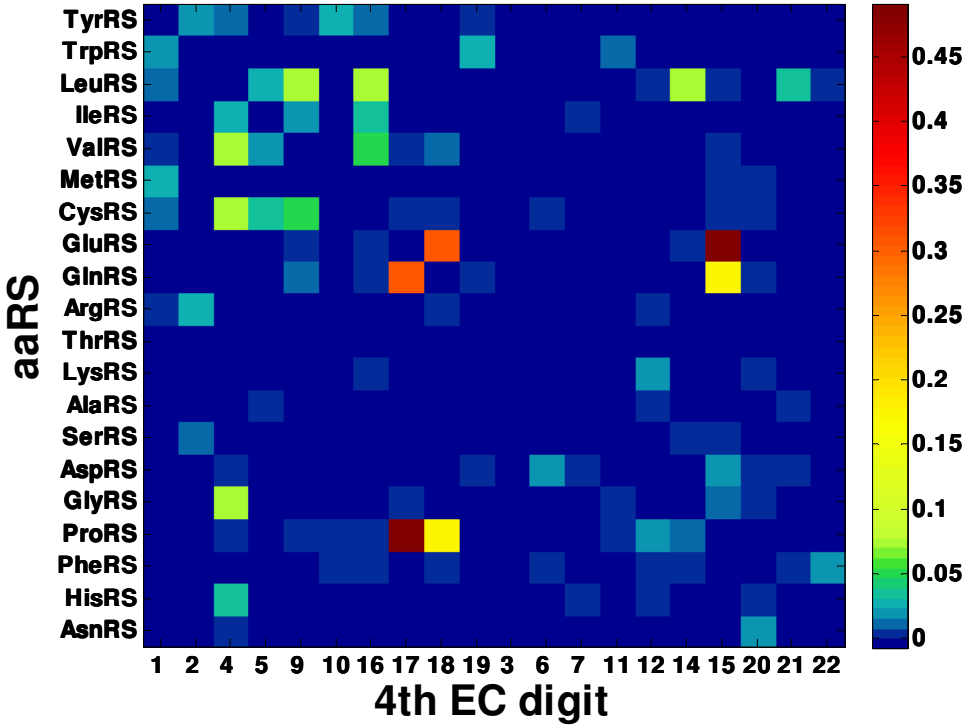


Figure S3. Heat map of Pearson cross-correlations of different aaRSs according to their shared CPs in Eukarya. Self correlations were left out for the purpose of clearer presentation.

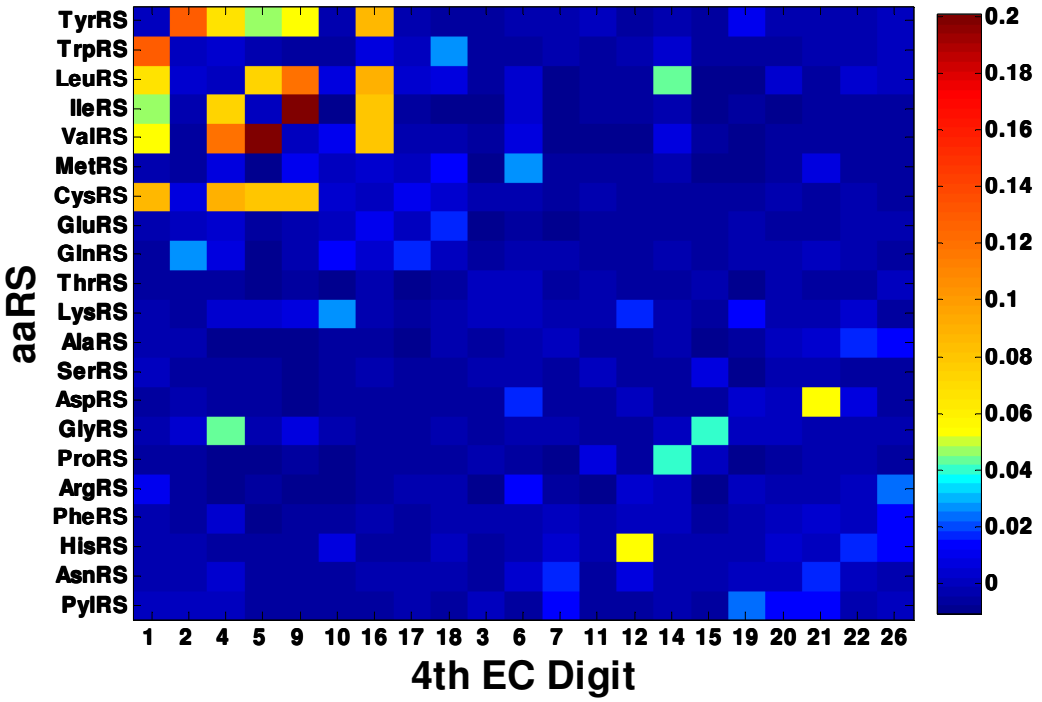


Figure S4. Heat map of Pearson cross-correlations of different aaRSs according to their shared CPs in Archaea. Self correlations were left out for the purpose of clearer presentation.

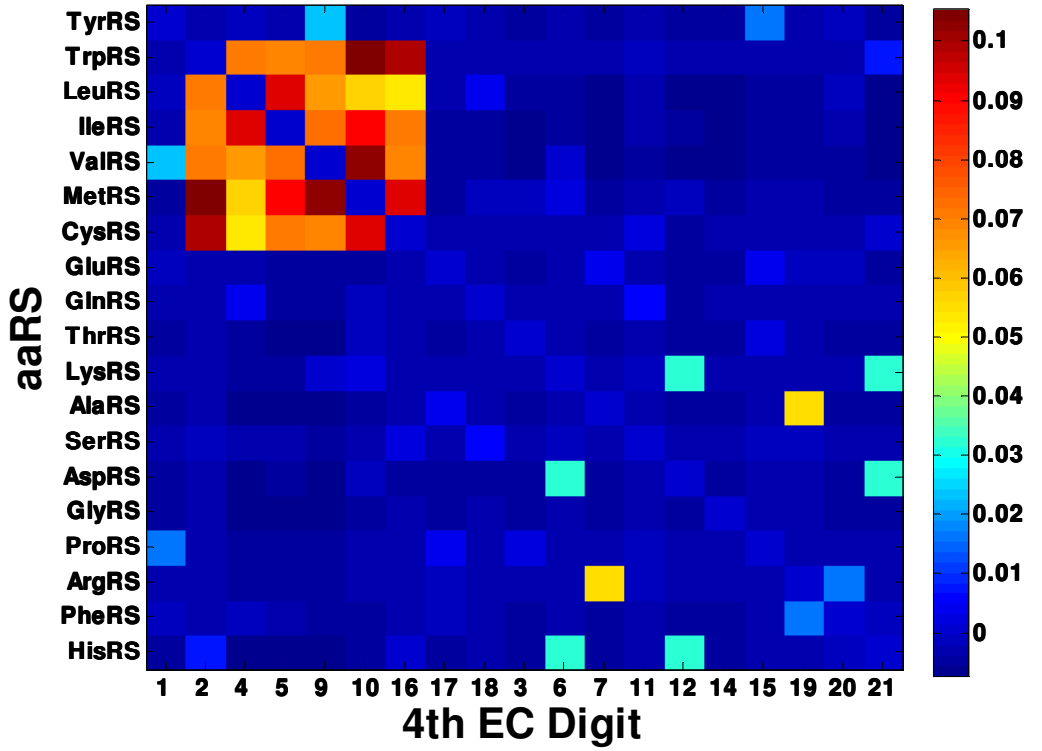


Figure S5. Heat map of Pearson cross-correlations of different aaRSs according to their shared CPs in Mitochondria. Self correlations were left out for the purpose of clearer presentation.

7.6.2 Evolutionary Aspects of CPs

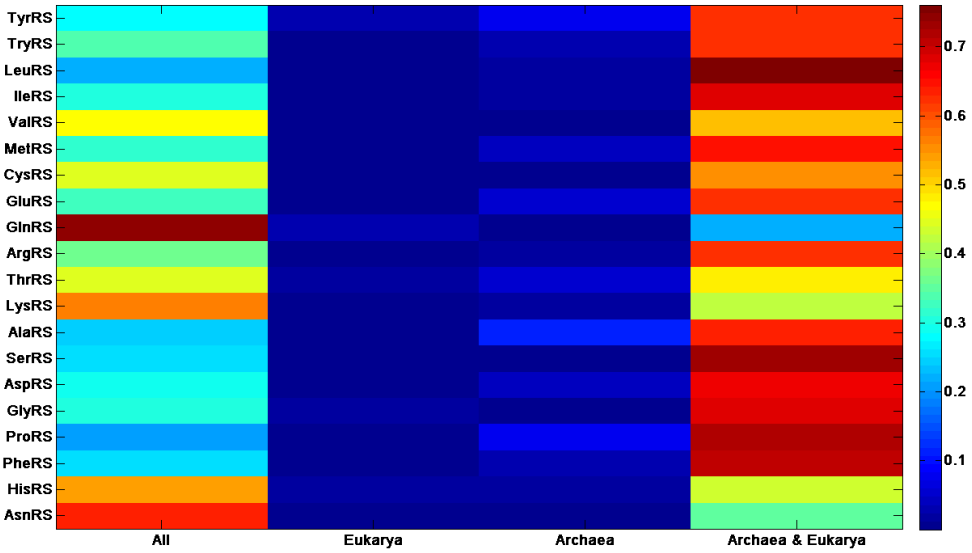


Figure S6. Distribution of different aaRS according to the CPs appearing in all 3 kingdoms together (All), in Bacteria and Eukarya together (excluding Archaea) and in Eukarya and Archaea kingdoms exclusively.

Chapter 8

Summary

This thesis presents methods and algorithms for unsupervised extraction of structures in biological data sets. It is divided into two distinct parts.

The first part introduces a novel unsupervised feature selection algorithm, which is later on developed into a complete framework, offering users a web tool for extracting features from various data sets.

The second part introduces the concept of Common Peptides and its application to vertebrate olfactory receptors and to aminoacyl tRNA synthetases.

In essence, both parts deal with unsupervised extraction of relevant features from very different types of data. However, since this task is complex in nature, one needs to match and tailor different solutions for different data types. In this thesis we have shown solutions that perform best for two commonly used biological data types – gene and microRNA expression arrays and protein sequences.

Since each chapter contains its own summary, we bring here some general insights gained from the development of methods described in this thesis:

1. No *one size fits all* solution.

   Each data type needs inspection and analysis in order to fit the best solution.

   Furthermore, in many cases an array of processing techniques and availability of tools that can be tailored together are the path by which interesting patterns emerge from the data.

2. In many cases, the data dictates the solution.

   As has been emphasized in the unsupervised motto of this thesis, the best way to let hidden patterns in the data expose themselves is to analyze the data according to its internal structure instead of fitting predefined models to it.

3. Unsupervised extraction is a relatively uncharted land but should be given more emphasis in the future.

   As supervised extraction of features is gaining much attention and effort, one has to realize that current inflation in biological data gives rise to problems in which neither the question nor the expected answer are pre-defined. In such scenarios, unsupervised analysis of the data may serve as a first step to understand the questions that may be asked and hopefully answered using the given data.

In conclusion, this thesis presents methods that attempt to "separate the wheat from the chaff" in biological data using unsupervised approach. These methods enrich the small repertoire of unsupervised data analysis and may benefit the study of complex biological systems that many researchers are attempting to decipher.

תקציר

כמויות וסוגי המידע במדעי הטבע גדלים במהירות. על מנת להתמודד עם כמויות גדולות אלו, קיים צורך בשיטות להפשטת המידע, התמרתו ומציאת מבנים בו. קיום אפשרי של קשרים מורכבים מצריך גם שילוב בין מקורות מידע שונים. ביולוגיה מהווה דוגמא טובה לתחום בו קיים מגוון רחב של מקורות מידע.

מציאת מבנים במקור מידע נעשית לרוב בצורה מונחית, על ידי התאמה בין הנתונים לידע קודם (לדוגמא התאמת קבוצות לתיוג ידוע). מציאת מבנים בצורה לא מונחית, לעומת זאת, חוקרת ומזהה מבנים הטבועים במקור המידע מבלי כל ידע קודם נוסף. כמות הנתונים הביולוגים, רובם ללא ידע רב קודם, מקשה על מיצוי מידע בעל משמעות. עובדה זו מספקת את הבסיס לחקר נתונים בצורה לא מונחית ומציאת מבנים במקורות מידע ביולוגיים.

תזה זו מתמקדת בשני נושאים המשתמשים בצורה לא מונחית לניתוח מידע:

1. כלים ואלגוריתמים לכריית מידע בצורה לא מונחית.
2. ניתוח משפחות חלבונים תוך שימוש במציאת מוטיבים בצורה לא מונחית.

הנושא הראשון כולל שיטות לעיבוד מקדים וחקר נתונים, הקרויות בשם הכללי שיטות לכריית מידע. אנו מציגים שיטה חדשנית להקטנת מספר המימדים הנקראת סינון מאפיינים בצורה לא מונחית (– unsupervised feature filtering UFF). אנו מיישמים שיטה חדשנית זו על מגוון מקורות מידע ביולוגיים הכוללים נתוני מדידת ביטוי גנים בסרטן, איידס וצהבת נגיפית מסוג C וכן ביטוי מיקרו-רנא בסרטן. שימוש במאפיינים הנבחרים ע"י UFF לצורך מציאת קבוצות מאפשרת להפחית את הרעש ולמצוא קבוצות ברורות התואמות תיוג ידוע, כאשר תיוג כזה זמין. פרט לכך, קבוצות הגנים ומיקרו-רנא הנבחרות מועשרות הן במונחים בעלי שייכות והן במונחים מפתיעים. לגבי רוב הגנים והמיקרו-רנא המדורגים גבוה לפי שיטתינו קיים תיעוד הקושר אותם למחלות המסוימות בעוד לחלקם קשרים כאלה עדיין לא קיימים. מאפיינים נבחרים אילו יכולים להוות מידע בעל משמעות ביולוגית אמיתית.

הנושא השני עוסק במוטיבים רצפיים קשיחים שמוצא האלגוריתם הבלתי מונחה למציאת מוטיבים MEX. אנו פיתחנו שיטה ליצירת קבוצה בעלת משמעות של מוטיבים אילו המכונים "מוטיבים שכיחים" (CPs). קבוצה זו מהווה מסגרת המאפשרת מחקר של מגוון משפחות חלבונים כגון זיהוי קבוצות בתוך המשפחה, מציאת עקבות היסטוריות של אירועים אבולוציוניים וחשיפת דמיון רחוק בין חלבונים. יישמנו שיטה זו על קולטני ריח ומשפחות האנזימים אמינואציל-tRNA סינתטאז (aaRS). שימוש בשיטה זו על קולטני ריח אפשר לנו לעקוב אחרי אירועים אבולוציוניים בחולייתנים תוך גילוי הורדת יתירות בבני אדם ביחס ליונקים אחרים, אובדן המוני בשושלת הזוחלים והעופות וחשיפת ההיסטוריה של משפחות הקולטנים השונות. אנו אף מצביעים על מוטיבים המבדילים בין יצורים שוכני מים ויבשה ומזהים את מיקומן על רצף חומצות האמינו של הקולטן.

שימוש במוטיבים השכיחים על משפחות האנזימים אמינואציל-tRNA סינתטאז מאפשר זיהוי של התפלגות שונה של המשפחות בממלכות החיים השונות. שיטה זו גם מזהה מוטיבים המבדילים בין שתי מחלקות ידועות של משפחות האנזימים הכוללות איזורים בלתי ידועים עד כה על הרצף. מוטיבים נפוצים נוטים לחפוף איזורי קשירה וזירוז.

TEL AVIV UNIVERSITY אוניברסיטת תל-אביב
RAYMOND AND BEVERLY SACKLER FACULTY OF EXACT SCIENCES
SCHOOL OF PHYSICS AND ASTRONOMY
הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר
בית-הספר לפיזיקה ולאסטרונומיה

מציאה לא מונחית של מבנים במידע ביולוגי

תזה מוגשת לקראת תואר

דוקטור לפילוסופיה

על ידי

אסף גוטליב

הוגש לסנאט של אוניברסיטת תל אביב

ספטמבר 2009

עבודה זו התבצעה בהנחייתו של

פרופ' דוד הורן