FACHHOCHSCHULE WEIHENSTEPHAN

University of Applied Sciences

Fachbereich Biotechnologie

International Master in Informatics for Biotechnology

Master's Thesis

Development of the software application *ggc* to analyse and compare proteomes from different species

Author: Mildred Hofmann

Date: 10.01.2005

Supervisor:

Prof. Dr. Bernhard Haubold

Bioinformatik Zentrum - Fachhochschule Weihenstephan

D 85350 Freising

Ph: ++49-(0)8161/715274

Eidesstattliche Erklärung

gemäß § 23 Abs. 6 Prüfungsordnung

Ich erkläre hiermit an Eides statt, dass die vorliegende Arbeit von mir selbst und ohne fremde Hilfe verfasst und noch nicht anderweitig für Prüfungszwecke vorgelegt wurde.

Es wurden keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt. Wörtliche und sinngemäße Zitate sind als solche gekennzeichnet.

Freising, den 05.01.2005

Mildred Hofmann

Acknowledgements

I thank Prof. Dr. Benno Müller-Hill, Köln University, for suggesting the topic of this thesis. I also thank Prof. Dr. Bernhard Haubold, my supervisor at the university, for his numerousness ideas and the time he spent on this project. His continuous support and his corrections to this thesis were a great help and improved it a lot. In addition, I want to thank Prof. Dr. Frank Lesske for proof-reading my master's thesis.

Special thanks go to my boyfriend Andreas Kirschner for helping me out with his knowledge in the computer-science field and for always encouraging me in difficult moments.

Thanks very much to Sandra Buchner for the last year and the great time we spent together in Northern Ireland. She was the perfect fellow student to do this course and remains a very good friend.

I also thank Johannes Mandel for sharing the office with Sandra and me, thus we had much fun in the last months. I want to thank Kajetan Hundhammer and Michael Krappmann for their helpfullness in all questions.

I am very grateful to Dr. Daniel Berrar from the University of Ulster for correcting my English in this thesis and for helping Sandra and me with all our problems in Ulster.

Last, but not least, I want to thank my parents and my grandmother for giving me the opportunity to do this master course.

Table of Contents

1.Introduction

The completion of the genome sequences of three mammals (human [18], mouse [13] and rat [14]) and the availability of the almost finished chimpanzee genome provides the opportunity for investigating these organisms at the genome level. Exploring issues of protein evolution that are best addressed through the study of more closely related genomes are possible. One of the most powerful general approaches for unlocking the secrets of genomes is comparative genomics. Comparative genomics is the analysis and comparison of genomes from different species. The purpose is to get a better understanding of how species have evolved and to determine the function of genes. The size and distribution of their gene-families can be investigated to gain more information about organisms. Identifying lineage-specific differences by comparing the size of these families might reveal how physiological, anatomical and behavioural differences are reflected at the genome level [14]. For example, dozens of local gene family expansions have occurred in the mouse lineage. Most of these seem to involve genes related to olfaction, suggesting that these physiological systems have been the focus of extensive lineage-specific innovation in rodents [13]. Indeed there is a phenotypic difference between rodents and primates in their ability to smell.

1.1.Gene-families

Proteins that have a common ancestral gene, but not necessarily have the same function, are termed homologues. Homologous genes found in different organisms are called orthologues. Orthologues are most similar in sequence comparisons between two organisms and their gene products (proteins) are most similar in function. Paralogues are homologues in the same organism and arise evolutionarily mainly via gene duplication events. Since the orthologue provides the needed protein function, paralogous genes are free to mutate, yielding genes encoding proteins of new function. As a result, paralogous genes are often less similar in sequence comparisons to a homologue in another organism than the corresponding orthologous gene. All paralogues combined are referred to as a gene-family. Figure 1 clarifies these relationships by the means of an artificial example using the evolution of the olfactory receptor gene-family. Protein A in mouse and human had a common ancestral gene before the split of the primate and the rodent lineage. The function of protein A is very similar in both organisms. After the split, the olfactory receptor gene-family in mouse expanded and the paralogous proteins B, C and D evolved. The olfactory receptor gene-family in mouse is larger than in human with only two paralogous proteins A and Z. In sequence comparison studies no homolgy between Protein Z and the mouse-proteins can be found.

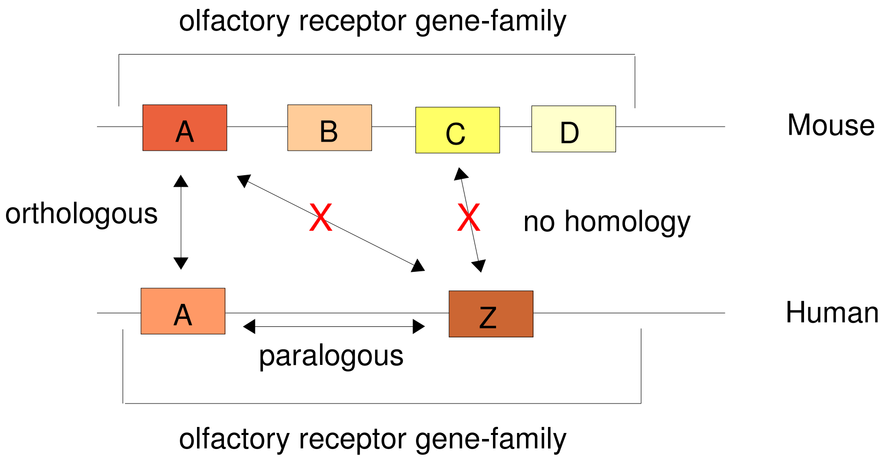

olfactory receptor gene-family
A
B
C
D
Mouse
orthologous
X
X
no homology
A
Z
Human
paralogous
olfactory receptor gene-family

Fig.1 Homology: orthologous and paralogous genes
The DNA-strands of human and mouse are illustrated with a line, the proteins A, B, C, D and Z with a rectangle. Protein A is found in mouse and human and has a very similar function (-> orthologous). Protein Z is paralogous to Protein A in human, but no homology is found to the proteins in mouse. In the mouse proteome Protein B, C and D are paralogues of Protein A

A very common similarity search tool that finds homologous sequences is the BLAST-software (Basic Local Alignment Search Tool) [1]. BLAST implements a rapid pair-wise comparison of a query sequence against a database, using a heuristic method. Each comparison is given a score reflecting the degree of similarity between the query and the sequence being compared. The higher the score, the greater the degree of similarity. By the means of this score the expectation value (E value) is calculated, which reflects the expected number of hits by chance.

The conservation of sequences is critical only in certain regions such as functional domains, domains required for the structural integrity of the protein or for binding of ligands. They remain highly conserved at specific positions which have very narrow requirements in the physiochemical properties. Conserved sequences within such domains are described as motifs and consist of amino acids conserved at specific positions interspersed with degenerate sequence. Such motifs are identified and represented by a range of techniques.

They are held in so called secondary protein databases, applicable for protein classifications. The information derived in this way usually takes the form of sequence patterns or the information is represented as profiles representing all member-sequences or statistical models (HMMs) of the members. The databases consist of libraries of such patterns, profiles or models against which a sequence can be compared. Examples of secondary databases are PROSITE (patterns, profiles) [17] and PFAM (HMMs) [3].

1.2.Evolutionary relationships

Over an evolutionary timescale organisms diverge and as the evolutionary distance increases, the sequence of any given protein changes due to mutations that occur on the nucleotide level. These changes can be subject to evolutionary pressure, but sequences change in the absence of any evolutionary pressure aswell. This is known as neutral evolution or drift. The number of mutations in a protein sequence that occur in a certain time interval can be estimated if the mutation rate per year is known.

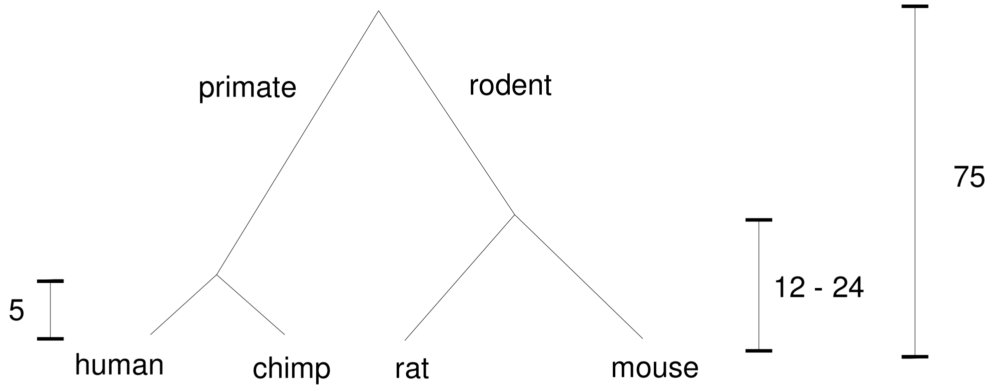


Fig.2 Simplified schema of a phylogeny representing the relationship between the primate and the rodent lineage
Time scales are in million years.

Figure 2 shows a simplified phylogeny of human, chimpanzee, mouse and rat. The rodent and the primate linage split approximately 75 million years ago. The last common ancestor of rat and mouse lived about 12-24 million years ago [14]. Human and chimpanzee split very recently, around 5 million years ago [16].

Due to the Rat Sequencing Project Consortium [14], the primate and the rodent genomes encode similar numbers of genes and the majority has persisted without deletion or duplication since the last common ancestor of the two lineages. For example, 80 % percent of mouse proteins seem to have strict 1:1 orthologues in the human genome [13]. But the remainder are of special interest because many belong to gene-families that have undergone differential expansion in at least one of the two genomes. More genomic changes occurred in the rodent than the primate lineages [14].

1.3.Conceptual formulation

The objective of this project was to implement a software that discovers gene families, evolved in a specified time interval, in the proteome of an organism and provides an overview about lineage-specific expansions by comparing the gene-family-sizes of two organisms. The difference to already existing software, like the PFAM-software [3], should be that no large, early in time evolved families are considered. Only families that evolved in the time interval of interest are grouped and compared.

The last task was the application of the implemented software with the aim to discover differences at the proteome-level between the primate and the rodent lineage, or inner-lineage differences between human-chimpanzee and mouse-rat.

2.Material and Methods

All URLs are listed in a seperated table in the appendix (-> 9.1).

2.1.Protein data sets

2.1.1.Source

The protein data sets analysed in this study were obtained from the following websites:

Mycoplasma genitalium Proteome

Using the EMBL-Sequence retrieval system SRS query form, the database UNIPROT_SPROT was searched for all protein sequences of *Mycoplasma genitalium* and saved in fasta format. Additionally an annotation file had to be saved with the UNIPROT_SPROT view because the fasta files contained just the protein names, without any annotations.

Mammalian Proteomes

The mammalian proteomes were downloaded from the NCBI and the ENSEMBL databse because the data-sets were not identical.

The NCBI fasta files of the proteomes (protein.fa.gz) of *Homo sapiens* (Build No. 35),*Mus musculus* (Build No. 35) and *Rattus norvegicus* (Build No. 1) were downloaded by ftp. These data-sets contained annotations of the protein sequences. Additionally the Genbank summary files (*.gbs) were downloaded to obtain information about the localisation of the proteins in the genomes.

The proteomes of *Homo sapiens* (Version: NCBI35, November 2004)*, Mus musculus* (Version: NCBIm33, July 2004)*, Rattus norvegicus* (Version: RGSC3.1, July 2004) and *Pan troglodytes* (Version: CHIMP1, May 2004) were downloaded by ftp from the ENSEMBL-server in fasta format (*aug.pep.known.fa.gz). Additionally, the genomes were downloaded as EMBL flatfiles for the localisation of the proteins in the genomes (*.dat.gz).

2.1.1.Cleaning of the data-sets

In the mammalian data-sets from NCBI and ENSEMBL occasionally more than one accession number was assigned to the same protein. Such a situation can be detected by comparing their sequence coordinates. A list of accession numbers with unique sequence coordinates was generated using the 'localisation files' and with this list a non-redundant fasta-file could have been constructed.

This procedure was carried out for the data sets obtained from both databases (NCBI and ENSEMBL) because both the number of proteins (Tab. 1) and the protein-sequences differed between the two data-sets. One reason for these discrepancies might be that many genes are predicted by automated computational analysis. Occasionally the various algorithms predict different genes or splicing variants.

	ENSEMBL		**NCBI**	
	all known proteins	**non-redundant**	**all known proteins**	**non-redundant**
human	28374	25578	27960	26684
mouse	24546	23410	26180	24728
rat	*5922*	-	21178	20080
chimp	28590	24376	-	-

Tab.1 Comparing the proteome sizes of the data-sets maintained from ENSEMBL and NCBI (before and after the cleaning step)
The rat ENSEMBL data-set (italic) could not be analysed, because of the limited no. of proteins in the data-set.
The NCBI database did not contain a chimp proteome data-set (October 2004)

2.2.Protein classification databases

The function of the generated gene-families was determined by searching the protein classification databases PROSITE [17] and PFAM [3].

PROSITE is based on a mixture of regular expressions (patterns) and profiles and consists of two flat files: *prosite.dat* and *prosite.doc*. This database can be used to rapidly classify a protein sequence. If no annotation was found with that method, the next step was to scan the PFAM libraries *Pfam_ls* and *Pfam_fs*. PFAM contains multiple protein alignments and profile-HMMs (Hidden Markov Models) of families. As the PFAM protein classification is based on HMMs, it is more sensitive but much slower than PROSITE. For example a protein sequence with 313 amino acids was searched against both databases. The PROSITE database search lasted 24 seconds; the PFAM search lasted with 3 min 12 sec approximately 6 times longer.

2.3.Software

The aim of this study was to develop a software that groups proteins of an organism into gene-families and compares the size of these gene-families between different organisms with the assistance of publicly available software like BLAST or HMMER. I developed a number of software modules and these were subsequently merged with the public software into two major programs.

2.3.1.Public available software-tools

Sequence-alignment software: `blastall` and `blastpgp`

For the local sequence alignments of the amino acid sequences the software-tools `blastall` [1] and `blastpgp` [2] were used. `blastpgp` performs gapped BLAST searches and can be used to perform iterative searches in position-specific iterated BLAST (PSI-BLAST). Sequences found in one round of searching are used to build a position specific score model for the next round of searching. The higher the similarity between two sequences the higher the score. Hits two more distant relatives receive a lower score. The profile is used to perform subsequent searches and the results of each iteration are used to refine the profile. The software `formatdb` formats protein sequences as databases so that they can be searched efficiently by `blastall` and `blastpgp`.

Protein classification tools

The first step to determine the function of a gene-family was scanning the database *prosite.dat* with the program `ps_scan.pl` [7]. Frequently matching (unspecific) patterns and profiles were skipped.

The next step was to search the libraries *Pfam_ls* and *Pfam_fs* with the program `hmmpfam` for annotations. This program is part of the HMMER software package [5].

2.3.2.Developed sub-programs

All software-applications generated in the context of this project were programmed in the computer language PERL [19]. This chapter gives a survey of the functionality of the developed programs. The majority of these programs accept user-defined input parameters. Examples of parameter-settings are given in chapter 2.4 and 2.5.

Program: `parser`

The program was developed to parse a `blastall-blastp` output file.

According to Gu et al. [9] it requires a more rigorous analysis than the expectation value returned by BLAST as a sole criterion to decide whether two proteins are homologous or not. The E-value of an alignment is dependent on the length of the sequences. If the E-value threshold is set too low, shorter sequences are not listed and with a high E-value too many proteins would be classified as homologous. Thus our criteria for two proteins to be homologous are alignable region (percent-coverage) and percent-identity.

The coverage is calculated by summing up the alignable regions (all aligned amino acids) of both proteins. The shorter alignable length is chosen and divided by the length of the longer protein [9].

The identity is obtained by dividing the sum of identical residues by the length of all aligned residues. If the alignable region is longer than 150 amino acids the calculated identity (I) should lie over the user-defined threshold, otherwise over the HSSP-curve (formula given below) [15].

Formula of the HSSP-curve: $I \geq 0.01n + 4.8L^{-0.32(1 + \exp(-L/1000))}$

$n = I - 24$; L = alignable length

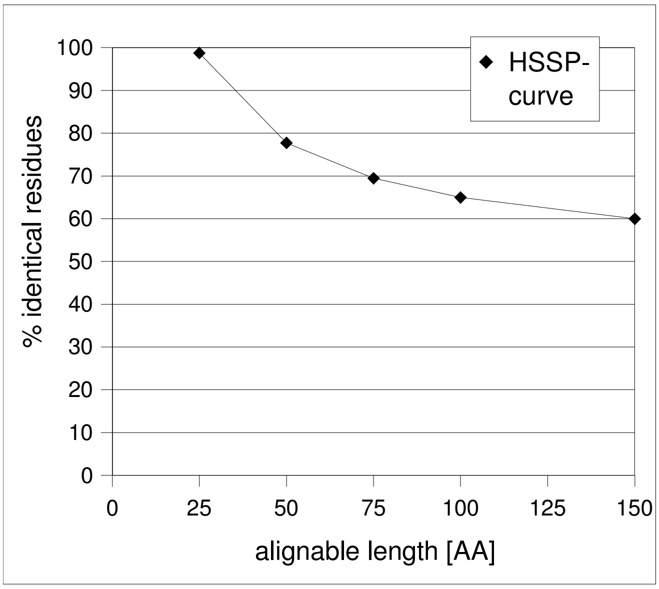


Fig.3 Graph of the HSSP-curve using an Identity threshold of I = 60 %

The graph in Figure 3 demonstrates the usage of the HSSP-curve, if the alignable length is shorter than 150 AA, the identity between the two protein sequences must be higher than the value in the HSSP-curve otherwise over an identity threshold of I = 60 %.

The HSSP-curve was derived from an empirical study which suggested that a higher I-value was needed for shorter proteins. The formula is continuous at L = 150 with n = I - 24 [9].

The program `parser` returns the accession numbers of the matched protein pairs, their coverage and identity.

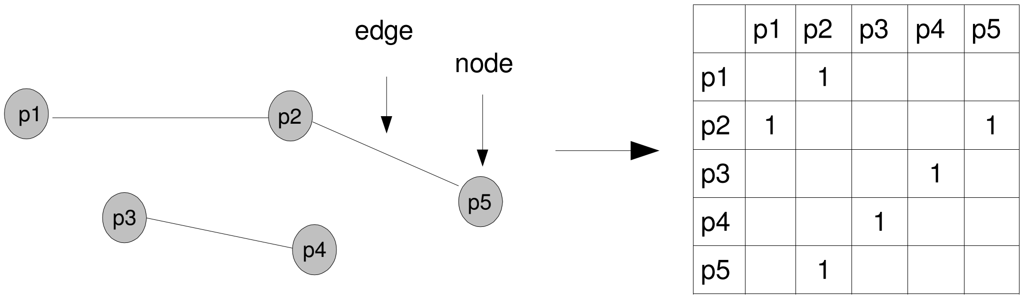
Depending on the sequence database used, NCBI or ENSEMBL, two versions of `parser` are available (`parser_ncbi`, `parser_embl`)

Program: `group`

Using the depth first search algorithm (DFS) [10] the program groups the protein pairs derived from `parser` into clusters. This algorithm accounts for the transitivity of homology, i.e., if protein A is homologous to protein B and protein B is homologous to protein C, using the DFS, protein A, B and C are grouped together regardless of whether BLAST finds a significant homology between protein A and protein C.

The program was implemented as follows:

An (m x m) matrix, where m is the number of all proteins in the proteome is constructed. If a protein-pair passed the identity and coverage criteria implemented in `parser,` the matrix is set to one at both positions (matrix[p1][p2] = 1 and matrix[p2][p1] = 1). Filling in the whole matrix is necessary if the DFS is used.



	p1	p2	p3	p4	p5
p1		1			
p2	1				1
p3				1	
p4			1		
p5		1			

Fig.4 Example graph and its corresponding representation, which can be traversed by the DFS-algorithm

The DFS-algorithm is the simplest approach to traverse a graph. A small example graph and its matrix representation are pictured in Figure 4 to clarify the technical terms. The algorithm starts by looking at the first node (prot1) in the x-direction (Fig. 4). This node is marked as visited and now all nodes (prot1..prot5) in the y-direction are looked at. If an edge is found (e.g. matrix[p1][p2] = 1) the edge is marked as visited (matrix[p1][p2] = 0) and if the node (prot2) in the y-direction was not already visited, the algorithm is called recursively until the edge (matrix[p2][p5] = 1) is found and marked as visited. In this way the DFS algorithm traverses the matrix and finds all members of a group. For each group a new DFS search must be started.

The program returns:

- The total number of grouped proteins and the total number of generated groups
- A list of all groups containing:
  No. of group members, an assigned group number, annotation of first protein (NCBI) and the accession numbers of all proteins.

There are also two versions of `group` available (`group_ncbi`, `group_embl`).

Program: `parser-prosite`

The first sequences of all groups of proteins are searched against the PROSITE database *prosite.dat*. This program parses the output file and returns a table containing the group no. and the PROSITE-annotation.

Programs: `getAllSeq`, `getOneSeq`, `getGroupSeq`, `oneSeq`

The first two programs generate sequence files in fasta format of one protein per group (`getOneSeq`), or of all grouped proteins (`getAllSeq`) for both formats (ENSEMBL, NCBI).

`getGroupSeq` returns the sequences of all proteins in a defined group and `oneSeq` the sequence of a single protein.

Program: `compare`

In order to compare protein groups of two organisms, another BLAST search is necessary. This program reports `blastpgp` matches of a group of organism1 with groups of organism2, if at least 20 % of the group members of organism2 were matched. The decision criterion for two proteins to be homologous is the expectation value of the `blastpgp` search.

`compare` reports the group number of the query sequence, number of group members, matched group numbers, their number of members and counts all hits to the sequence database of the other organism, the matched groups and the matched sequences per database group.

Program: `parser-group`

Another approach to match groups of two organisms was to perform a `blastall-blastp` search of all grouped proteins of organism1 against a database of all grouped proteins of organism2. This program finds matches between groups by parsing the `blastall` output, using coverage and identity (-> `parser`) as decision criteria. A group is said to be homologous if at least one protein-pair is found with these criteria.

The program returns a list of all groups (group no.) of organism1 with either the group no. of the matched groups of organism2 or the notice that no significant hit was found.

Program: `cluster`

This program clusters the matched groups derived from `compare (cluster)` or `parser-group (cluster-group)` using the DFS. The clustering step is necessary because a group of organism1 might match more than one group in the other organism and these groups can be matched by another group of organism1 additionally (Fig. 5). These relationships had to be taken into account for comparing the size of gene-families. In the example shown in Figure 5 the members of the groups g1 and g3 (organism1) are counted and compared with the total number of proteins in the groups G2, G3 and G4 of organism2.

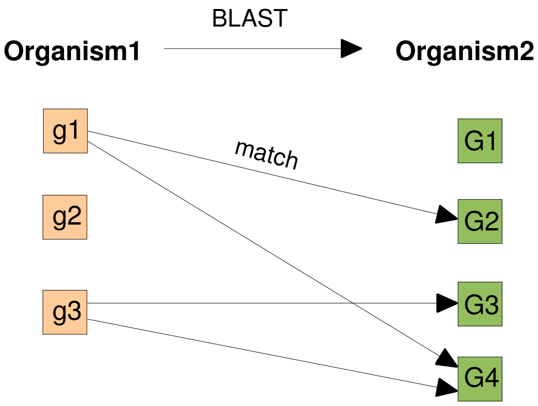


Fig.5 Example of possible BLAST matches between gene-families of two organisms
BLAST-direction: Organism1 -> Organism2

A cluster is printed out, if the summed up number of group members of the organism with more members lies over a specified threshold. Another decision criterion is the ratio of the number of members of both organisms in the cluster. The groups with no significant hit to a group of organism2 are printed out if their number of members is larger than specified.

`cluster` returns four files:

- *org1_org2.cluster:*

  This file contains a list of all groups with no significant match and a second list of the located clusters with the group no. of all clustered groups, their number of members (proteins) and if available their PROSITE-annotation.
- *org1_noMatch.fasta:*

  A fasta file with the sequences of all proteins (`cluster-group`) or one protein per group (`cluster`) of the groups that did not match a group of organism2 is generated for another `blastall-blastp` search against a database of all proteins of organism2.
- *org1_pfam.fasta, org2_pfam.fasta:*

  These sequence files are generated for a search against the PFAM database to get annotations for groups without a PROSITE annotation. Only the sequence of the first protein in the group is chosen because the PFAM database search takes very long and the sequences of the group-members are very similar.

Program: `parser_noMatch`

The BLAST output file (*org1_noMatch.fasta* against a database of all proteins of organism2) is parsed and all the matches of a group against this database are counted.

Program: `cluster2`

This program adds the PFAM-annotations and the information about BLAST hits to the clustering results of `cluster (cluster2)` and `cluster_group (cluster2_group).`

The final results are printed to two files (*org1_org2_result, org1_org2.da*t):

- *org1_org2_result* gives an overview of the unmatched groups and clusters and their functions in the form of two tables:

  Table1 (unmatched groups):

  #No.|annotation -source|no. of proteins |matches with db of all seqs of org2

  Table2 (cluster):

  #No.|annotation -source|no. of groups with this annotation (total no. of groups)

|total protein (org1)| total proteins (org2) |Ratio (larger/smaller)

|Organism with more proteins in the cluster

- *org1_org2.dat* prints the details (accession numbers, etc) of the picked groups for future research:

  Table1 (unmatched groups):

  #No.|group number|protein accession number |organism

  Table2 (cluster):

  #No.|group number|protein accession number |organism

2.3.3.Major application software for grouping and comparing gene-families

Program: `grouping`

The program `grouping` combines all the steps that are necessary to generate the protein-groups (gene-families) of an organism (Fig.6). The grouping of the amino acid sequences can be carried out with fasta files obtained from ENSEMBL or NCBI. The parameters, `blastall` expectation value, coverage and identity (-> `parser`), can be specified by the user.

The program calls:

- `formatdb:` generates a database of the sequence-file for the BLAST search
- `blastall-blastp:` used for all against all comparison

  (filtering parameter is set to false: -F F, if parts of the sequences are masked by BLAST, the observed identity of two sequences is calculated too low)
- `parser`: parses the BLAST output file
- `group`: groups proteins into gene-families
- `get_allSeq`: generates fasta-file of all grouped proteins
- `get_oneSeq`: generates fasta-file of one protein per group
- `ps_scan.pl`: scans the PROSITE database *prosite.dat* for group annotations
- `parser_prosite`: generates table with PROSITE annotations

The program returns two result files, the grouping results (*orgCov_Ident.group*) and a list of group annotations (*orgCov_Ident_prosite.table*)

- *orgCov_Ident.group* contains: Total no. of grouped proteins and total no. of groups

  No. of proteins in group (0..x): List of protein accession no.
- *orgCov_Ident_prosite.table* contains a table of all groups with their PROSITE description

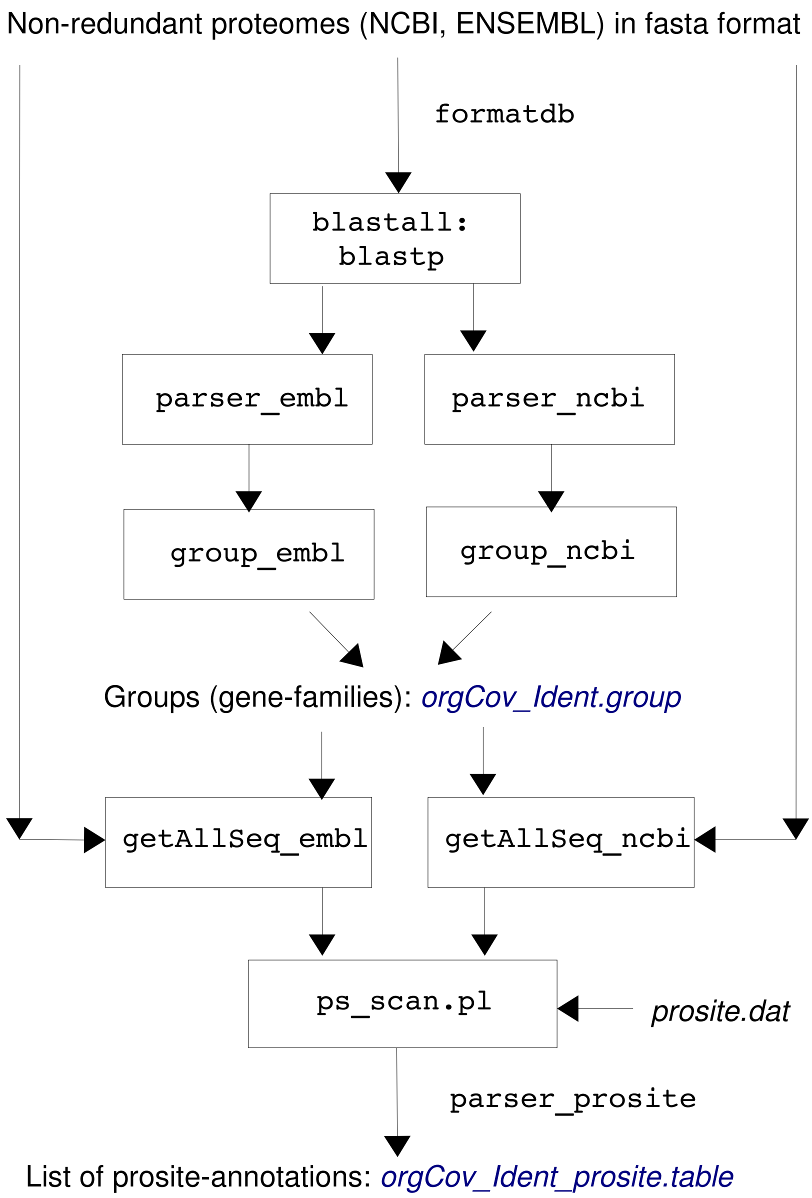

Non-redundant proteomes (NCBI, ENSEMBL) in fasta format
formatdb
blastall:
blastp
parser_embl
parser_ncbi
group_embl
group_ncbi
Groups (gene-families): orgCov_Ident.group
getAllSeq_embl
getAllSeq_ncbi
ps_scan.pl
prosite.dat
parser_prosite
List of prosite-annotations: orgCov_Ident_prosite.table

Fig.6 Flow chart of the processing steps implemented in the program `grouping`
Result files are indicated in blue

Program: `comparing`

This software integrates all steps necessary to compare the size of gene-families generated with `grouping`. The aim is to find differences in the size of gene-families of two organisms or gene families with no matches to a group of the other organism. The matching step (BLAST search) has to be carried out in two directions to confirm the clustering results and to find groups with no matches to the other organism. Thus all steps listed below are carried out twice. The major processing steps of the first matching direction are visualised in a flow chart (Fig.7) to clarify the structure of the software.

`blastall` directions: Organism1 (query sequences) -> Organism2 (database)

Organism2 (query sequences) -> Organism1 (database)

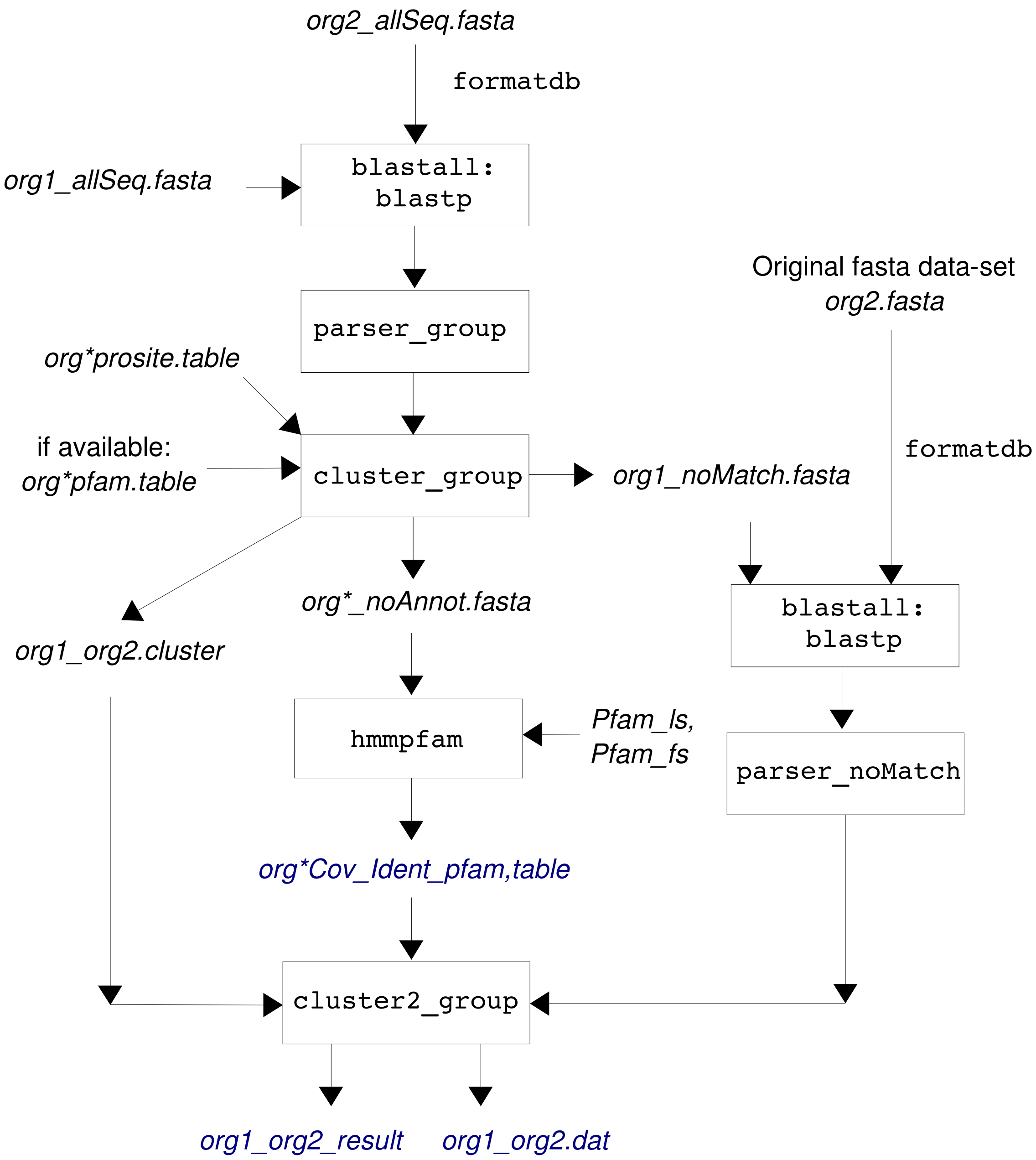


Fig.7 Flow chart of the major processing steps implemented in the program `comparing`
First `blastall-blastp` direction (organism1 -> organism2) in the first round (R = 1, explained in the text), Result files are indicated in blue

The variable parameters are identity and coverage (-> `parser`) in the parsing step and minimum no. of proteins in groups with no matches against database, minimum no. of total proteins of organism with more proteins in the cluster and minimum ratio: larger/smaller (-> `cluster`) in the clustering step.

Depending on the size of the input sequence, the program may run for several hours. Very time intensive steps are the BLAST and PFAM databases searches, but the BLAST search needs to be carried out only once. The program can be executed in another mode if in a second run (R ≥ 2) only the parsing or clustering parameters are changed.

The program implements:

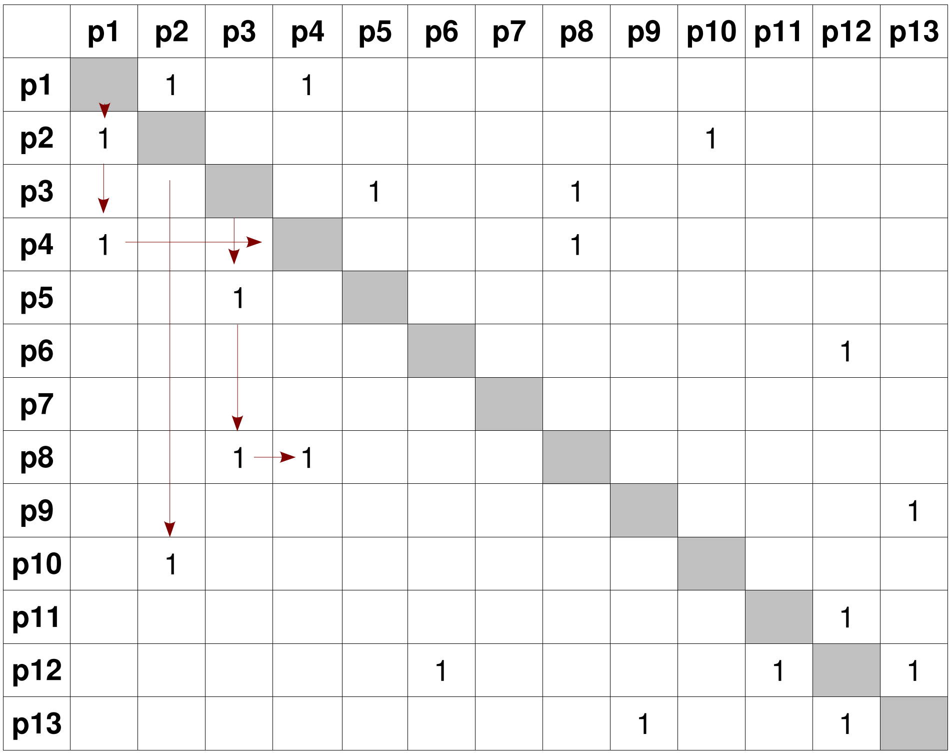
- `get_allSeq`: generates fasta-file of all grouped proteins
- `get_oneSeq`: generates fasta-file of one protein per group
- `formatdb`: generates database of all grouped sequences
- `blastall-blastp`: all grouped proteins against a db of all grouped proteins of the other organism (filtering parameter is set to false: -F F)
- `parser_group`: parses BLAST output files by printing matches between groups of both organisms as pairs
- `cluster_group`: finds cluster of grouped gene-families by means of the DFS
- `blastall-blastp`: proteins of groups with no significant match in the other BLAST search are compared against a database of all proteins of the other organism (filtering parameter is set to false: -F F)
- `parser_noMatch`: matches of a group with database of all proteins are counted
- `hmmpfam`: searches PFAM databases *Pfam-ls* and *Pfam-fs* for group annotations
- `cluster2_group`: prints result and data files

The program returns:

- Two result files (*org1_org2_result, org2_org1_result*) containing a table of unmatched groups and a second table of the clusters that fulfilled all conditions (-> `cluster`)
- Two data files with the accession numbers and group no. of the proteins (*org1_org2.dat, org2_org1.dat*) with a table of the unmatched groups and a table of the clusters.
- Two PFAM-annotation files (*org1Cov_Ident_pfam.table, org2Cov_Ident.table*) with a list containing PFAM-models and the PFAM-descriptions of the groups. These annotation files can be reused if the coverage and identity thresholds are not changed.

2.4. Test-runs to check the reliability of `parser` and `group`

2.4.1. Test with a simple example



	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13
p1		1		1									
p2	1									1			
p3					1			1					
p4	1							1					
p5			1										
p6												1	
p7													
p8			1	1									
p9													1
p10		1											
p11												1	
p12						1					1		1
p13									1			1	

Fig.8 Matrix of simple test example to check that `group` finds the correct cluster
p1-p13 are the proteins that should be grouped by the clustering algorithm.
A naïve approach is filling in half of the matrix and walking through the matrix as indicated with the red arrows.

The task was to implement an algorithm that accounts for the transitivity of homology, i.e. if protein A hits protein B and protein B hits protein C, all proteins A, B and C should be grouped together, regardless of whether A hits protein C or not. A naive algorithm that runs through the matrix as indicated by the arrows in Fig.8 groups p1, p4, p2 and p10 in the same cluster, but not p3, p8. To improve on this the depth-first-search algorithm (DFS) [10] was implemented in `group`. A test run to check the reliability of the program was carried out with this simple, made up example data.

2.4.2.Test with a small proteome (*Mycoplasma genitalium*)

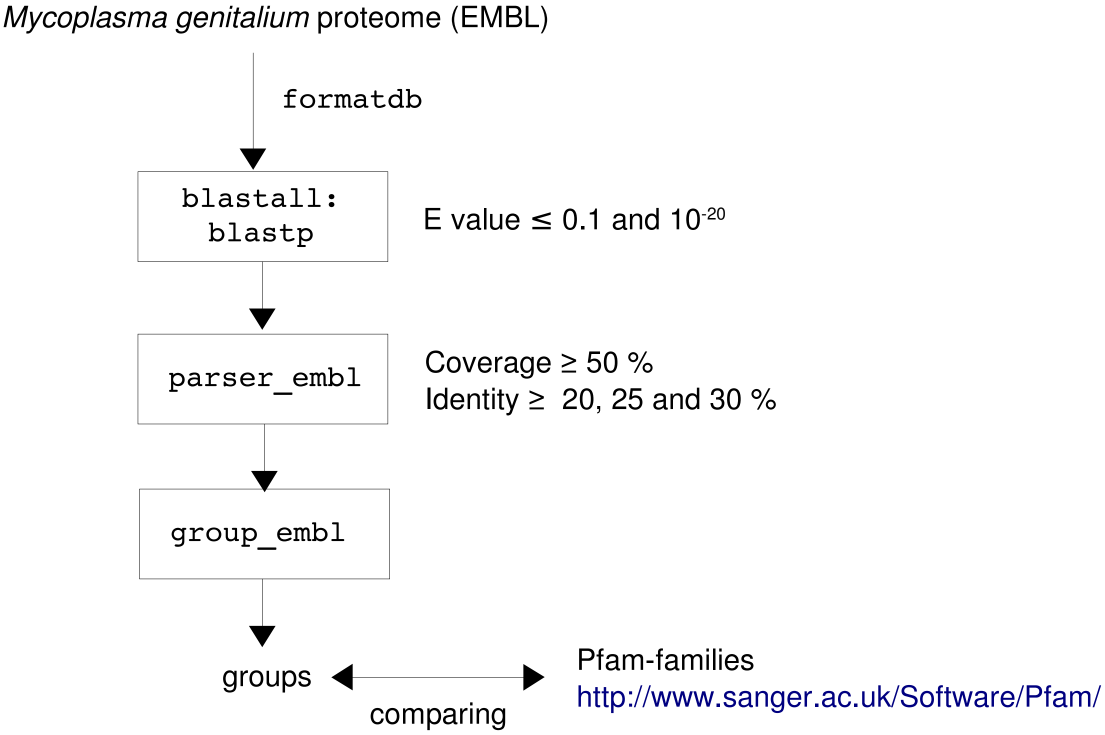

Mycoplasma genitalium proteome (EMBL)
formatdb
blastall:
blastp
E value ≤ 0.1 and 10^{-20}
parser_embl
Coverage ≥ 50 %
Identity ≥ 20, 25 and 30 %
group_embl
groups
comparing
Pfam-families
http://www.sanger.ac.uk/Software/Pfam/

Fig.9 Flow chart of the program test (`parser` and `group`) with the proteome of *Mycoplasma genitalium*

A test run with the smallest genome (*Mycoplasma genitalium* [6]) of a free-living organism was performed to verify that the two programs `parser` and `group` produce correct results. The genome was downloaded from EMBL with the sequence retrieval system (SRS), in order to work with the same protein names as the PFAM database. As shown in Figure 9 `blastp` was run with two different E values (E is the expected number of hits by chance) and `parser` with three different identity thresholds to determine the influence of these parameters. All parameters were set to low values to detect highly divergent gene-families as well. The last step was to compare each member of the groups obtained with the PFAM-families to ensure that the grouped gene-families are similar to the families generated with the Hidden-Markov-Models in the PFAM-database.

2.5. Analysis of the mammalian proteomes

2.5.1. Lengths distributions of the sequences in the data-sets

The lengths distributions of the amino acid sequences in the mammalian proteomes were investigated to get an overview about the quality of the data-sets. The human and mouse data-sets from both databases were analysed and compared with each other. The quality of the chimp (ENSEMBL) and the rat (NCBI) data-sets was tested additionally. The sequence-lengths were determined by counting the amino acids of each sequence. The lengths-frequency distribution was calculated from that data.

2.5.2. Generating human, chimp, mouse and rat gene families

Using the software `grouping,` the proteins of the human, chimp, mouse and rat proteome were clustered into gene-families. The parsing parameters identity and coverage were set to more stringent values than for the grouping of the proteome of *Mycoplasma genitalium*, because only younger gene-families should be located. The primate- and the rodent-linage for example split around 80 million years ago and it was the aim to find only families that evolved since that time. The last common ancestor of human and chimp lived about 5-7 million years ago [16], of rat and mouse at an outside estimate 25 million years ago [14]. The expected identity was determined by the means of the highest mutation rate in mammals [12]:

Primate-rodent: $3.06 * 10^{-9}$ [1/year] * 80 $*10^{6}$ [year] = 0.2448 ≈ 0.25
⇨ Identity: 75 %, Coverage: 80 %

Mouse-rat: $3.06 * 10^{-9}$ [1/year] * 25 $*10^{6}$ [year] = 0.0765 ≈ 0.08
⇨ Identity: 92 %, Coverage: 95 %

Human-chimp: $3.06 * 10^{-9}$ [1/year] * 7 $*10^{6}$ [year] = 0.0214 ≈ 0.02
⇨ Identity: 98 %, Coverage: 98 %

As it is not possible to calculate the expected coverage a sensible threshold was estimated. The E-value was set to $E \leq 10^{-20}$. The parameter-settings that were used to generate the gene-families are listed in Table 2. The proteins of both data-sets (NCBI and ENSEMBL) were clustered into gene-families.

Database	Organism	Coverage [%]	Identity [%]
NCBI	human	80	75
	mouse	80	75
		95	92
	rat	95	92
ENSEMBL	human	80	75
		98	98
	chimp	98	98
	mouse	80	75

Tab.2 List of the parameter-settings of `grouping` for the gene-clustering of the data-sets obtained from NCBI and ENSEMBL

2.5.3.Comparison of two approaches, matching gene-families between organisms

The matching step of the groups between two organisms was carried out with two different approaches, both based on a BLAST search.

Approach A: `blastall-blastp`

My first approach was to run a `blastall-blastp` search of all grouped sequences of organism1 against a database of all grouped sequences of organism2. The further processing steps of the data are pictured in Figure 7. A weak E-value should be set, because homology is determined on the basis of identity and coverage rather than the E-value.

Approach B: `blastpgp`

My second approach (Fig. 10) was to first generate a profile of a group and then to carry out profile search of the database of all grouped sequences of the other organism to find homologous groups. As the computation of multiple sequence alignments is very time and memory consuming, the group profiles were generated with the efficient PSI-BLAST software (`blastpgp`). The profiles for each group were generated with a `blastpgp` of one sequence of the group against all group members using a weak E-value. As shown in Figure 9, these profiles were then passed on as additional information to the second `blastpgp`, the matching step. In this approach two proteins were deemed homologous based on the expectation value of the `blastpgp` search. Thus the E-value had to be increased to match only close relatives. The major processing steps that are different to the other approach are described in Figure 10.

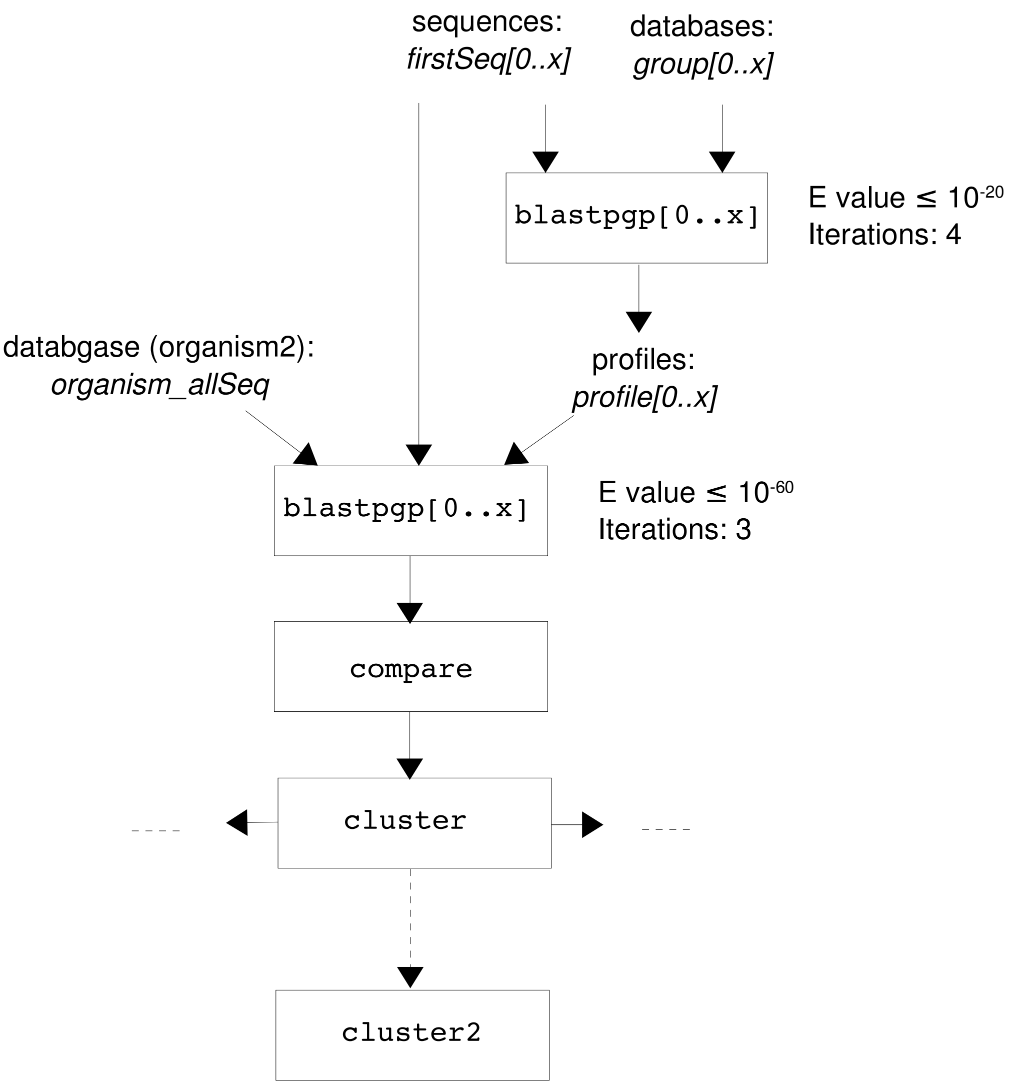


Fig.10 Flow chart of the main steps of the comparison of the protein groups of two proteomes using the PSI-BLAST program `blastpgp` for the group-matching step.

Using both approaches the gene-families of human and mouse (NCBI) were compared in both directions (query->database). The gene-families of the two organisms were generated with I ≥ 75 % and C ≥ 80 % (Tab. 2).

Parameter settings approach 1:

- `blastall-blastp`: expectation value E ≤ 10^{-20}
- `parser_group`: identity I ≥ 60 %, coverage C ≥ 80 %
  - ⇨ evolutionary distance between mouse and human is twice the distance for one 2 * 0.25 = 0.5 ⇨ expected difference = 50 %

BUT: ⇨ The probability that mutations hit the same amino acid position in a sequence more than once increases with the time interval. Two avoid matches between non-orthologous proteins the identity threshold was increased to I ≥ 60 %.

- `cluster_group`: (1) minimum no. of proteins in a group without matches: 10
  (2) minimum no. of clustered proteins of organism with more proteins in the cluster: 15
  (3) minimum ratio (larger/smaller): 1.5

Parameter settings approach 2:

- First `blastpgp`: expectation value $E \leq 10^{-20}$
- Second `blastpgp`: expectation value $E \leq 10^{-60}$
- `cluster_group`: (1) minimum no. of proteins in a group without matches: 10
  (2) minimum no. of clustered proteins of organism with more proteins in the cluster: 15
  (3) minimum ratio (larger/smaller): 1.5

2.5.4.Generating human gene-families with identical sequences

A very interesting part of the human proteome are protein families that evolved so recently that no amino acid mutations happened since that event. These very young families might reflect human-specific gene duplications. The `grouping` software was used to gain this information from both data-sets (NCBI and ENSEMBL).

Parameter settings:

- `blastall-blastp`: expectation value $E \leq 10^{-20}$
- `parser_ncbi` and `parser_embl`: identity I = 100 %, coverage C = 100 %

3.Results and Discussion

3.1.Results of the reliability test runs with `parser` and `group`

3.1.1.Test-run of `group` with simple example

A first examination of the `group` results was carried out with a simple example. Two input files (Fig.11) were constructed and the data was processed with `group_ncbi`. The appendant matrix is displayed in Figure 8 (-> 2.4.1).

```
p1|p2                       >p1|annotation1
p2|p4                       AAAAAAA
p2|p10                      >p2|annotation2
p5|p3                       BBBBBBB
p3|p8                       >p3|annotation3
p8|p4                       CCCCCCC
p6|p12                         .
p13|p9                         .
p12|p11                        .
p13|p12
```

Fig.11 Input files for `group_ncbi`: `parser`-output-file and sequence file in fasta format

```
7 members in group0 : annotation1
     p1,   p2,   p3,   p4,   p5,   p8,   p10,
5 members in group1 : annotation11
     p11,  p12,  p13,  p6,   p9,
```

Fig.12 Output of the program `group_ncbi`

The output displayed in Figure 12 demonstrates that the proteins are clustered correctly with `group`. This means that the program correctly accounts for the transitivity of homology. Furthermore, the output shows that if the input file was derived from the NCBI database (fasta files include annotations) the group-annotation is taken from the first group member.

3.1.2.Comparison of groups with PFAM-families of *Mycoplasma genitalium*

The reliability of the programs `parser` and `group` was examined with the smallest available genome, the genome of *Mycoplasma genitalium* [6]. The members of the generated groups were compared with the members of the PFAM-families. PFAM-families with only one member were excluded, as the generated groups consist of at least two proteins. Proteins that are members of two or more PFAM-families (Fig. 13) were only counted once.

```
PARE_MYCGE    30   174 PF02518 HATPase_c
PARE_MYCGE   220   391 PF00204 DNA_gyraseB
PARE_MYCGE   419   531 PF01751 Toprim
PARE_MYCGE   558   624 PF00986 DNA_gyraseB_C
```

FIg.13 PFAM-database output for protein PARE (*Mycoplasma genitalium*)
-> member of four PFAM-families

Identity and expectation value (E-value) were varied to examine the influence of these parameters. For this task the coverage was reduced to 50 %, because the 80 % requirement leads to more distant homolgues being missed [8] and the aim was to replicate the PFAM-families as closely as possible.

As expected, the results listed in Table 3 show that more proteins are grouped using a lower identity threshold. Not all generated groups could be matched with a PFAM-family, meaning that the proteins in these groups were not classified into a PFAM-family (118 proteins of the Mycoplasma proteome were not assigned to a PFAM family). Few of the generated groups were very similar to a PFAM-family but contained additional unclassified proteins. Only in one case (E ≤ 0.1, I ≥ 20 %) were members of different PFAM-families grouped together, which is undesired (Tab. 3, indicated in red).

The largest family in *Mycoplasma genitalium* with the PFAM-annotation ABC transporter, was divided in more than one group. Sometimes the arbitrary alignable-length limit prevents true homologues from being clustered in the same family [9]. However, this is preferable to grouping different PFAM-families in one cluster as a result of relaxing the decision criteria.

Additionally, the effect of varying the E-value ($E \leq 10^{-20}$ and E ≤ 0.1) was investigated. An advantage of a small E-value is that BLAST runs faster and the size of the output file is much smaller. But a loose similarity search criterion can improve the clustering [8] if a sensible identity and coverage threshold is chosen. The results (Tab. 3) were identical for both E-values with an identity threshold of I ≥ 30 %. Working with the weaker E-value caused a faster increase of the number of generated groups by reducing the identity threshold; but a more restrictive E-value prevented false classifications (indicated in red).

E-value	**E ≤ 10 $^{-20}$**			**E ≤ 0.1**		
Identity [%]	**I ≥ 20**	**I ≥ 25**	**I ≥ 30**	**I ≥ 20**	**I ≥ 25**	**I ≥ 30**
No. of groups (PFAM-families: 48)	**22**	**17**	**11**	**43**	**29**	**11**
Matches with PFAM-families	19	16	11	35	23	11
Groups with no match to a PFAM-family	3	1	0	8	6	0
No. of grouped proteins (proteins in PFAM-families: 151)	**54**	**43**	**30**	**125**	**72**	**30**
Members of PFAM-families	47	41	30	88	64	30
No members of a PFAM-family	7	2	0	29	8	0
False classified protein	0	0	0	8	0	0

Tab.3 Comparison of the *Mycoplasma genitalium* protein groups with PFAM-families
`blastall-blastp` parameter: E ≤ 0.1 and 10 $^{-20}$; `group` parameter: I ≥ 20, 25 and 30 %
48 PFAM-families with 151 members were selected (families with only one member and families that are part of another PFAM family were excluded);
368 out of 486 *Mycoplasma genitalium* proteins are members of a PFAM-family

The results (Tab.3) demonstrate that the generated groups do not disagree with the PFAM-families if the identity threshold and the expectation value are not set too low (no false classified proteins). The programs `parser` and `group` cannot locate all PFAM-families with the chosen decision criteria. This is not surprising, as the PFAM-database, based on Hidden-Markov-Models is more sensitive than BLAST. However, the purpose of this project was not the identification of very ancient paralogues. Instead the aim was to investigate the differences between recently evolved families of two organisms. With this objective in mind, the programs are running very reliably, because strong decision criteria can be employed. As there was no difference in the results for both E-values with I ≥ 30 % a more restrictive E-value was chosen to reduce the run time and disc-space requirement.

3.2.Results of the analysis of the mammalian proteomes

Having established the reliability of the developed programs, the next step was the analysis of the human, chimp, mouse and rat proteomes. As the size of the downloaded data-sets varied between both the different organisms and the two databases (Tab.1), further analyses of the data-sets was necessary.

3.2.1.Length distributions of the sequences in the data-sets

A first impression of the lengths of the protein sequences is gained by looking at the shortest and the longest sequence of the data-sets (Tab.4). The human and mouse proteomes, derived from the NCBI database, contain very large proteins. The human proteome derived from ENSEMBL contains a large protein similar in size to the largest protein in the NCBI data-set. The longest proteins of the mouse and chimp data-sets obtained from ENSEMBL are much shorter. The lengths of the shortest proteins are comparable, except for the shortest chimp-peptide with a length of only two amino acids.

Databank	Organism	Length of shortest protein [AA]	Length of longest protein [AA]
NCBI	Human	17	34,351
	Mouse	24	37,778
	Rat	24	12,338
ENSEMBL	Human	21	32,792
	Chimp	2	6,532
	Mouse	21	7,399

Tab.4 Length of shortest and longest amino-acid sequence of the mammalian data-sets

The lengths-frequency distributions were calculated for all mammalian data-sets and Figure 14 shows the distributions of the human (NCBI and ENSEMBL) and the chimp data-sets. The distributions of the two human data-sets are not identical, but very similar. This demonstrates that the discrepancies between the two databases are not excessive. The same observations were made with the mouse data-sets obtained from the two databases (data not shown).

The lengths-frequency distributions of the mouse and rat data sets were similar to the human distribution, but the rodent data-sets contained noticeably less proteins shorter than one hundred amino acids (data not shown). Remarkable are the wide differences between the chimp data-set and the other data-sets. More than 550 sequences are not more than 25 AA long and 71 of these entries are even shorter than ten AA.

These results confirm that there are some differences between the NCBI and ENSEMBL data-sets, but overall it seems that the data-sets are still very similar. Additionally, it is shown that the mouse and rat data-sets are comparable to the human data-sets. The reason for the significant difference of the chimp frequency distribution to all other distributions may be due to the fact that the completed chimpanzee genome was not published at the time these investigations were carried out.

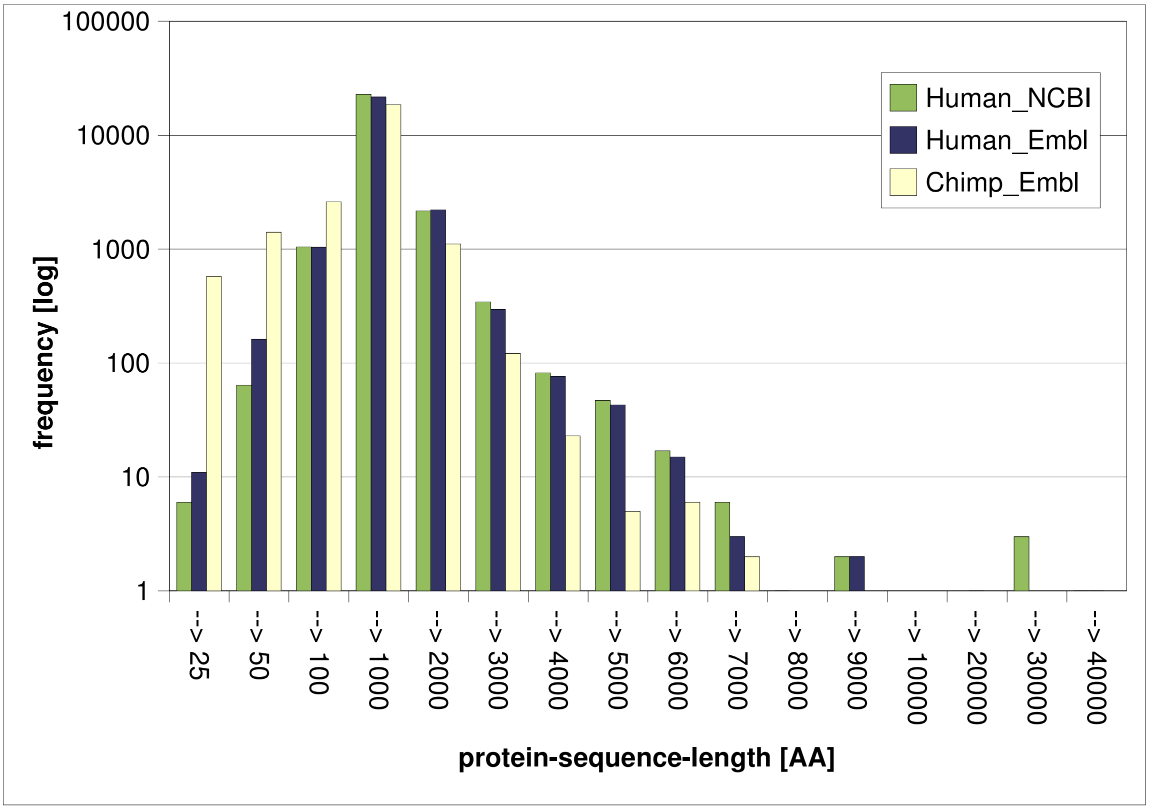


Fig.14 Logarithmic lengths-frequency distribution of the protein sequences in the human (ENSEMBL, NCBI) and the chimp (ENSEMBL) data-sets
The frequency scale is logarithmic and the lengths intervals increase for longer sequences.

3.2.2.Evaluation of the `group` output files

The content of the `group` output files were summerised to present the results in the form of two tables (Tab.5 and Tab.6). The evaluation shows that the number of generated groups differed by ca. 800 groups between the protein data-sets of NCBI and ENSEMBL (Tab.5). Using the ENSEMBL data-set, more gene-families were located and more proteins were grouped together. With both data-sets more groups could be found in the human proteome and the number of grouped proteins was higher as well, but larger protein-groups were found in the mouse proteome. The generated protein-groups reflect gene-families evolved since the split of the primate and rodent lineages about 75 million years ago.

	Human		Mouse	
Database	**NCBI**	**ENSEMBL**	**NCBI**	**ENSEMBL**
Proteome size [no. of sequences]	26684	25578	24728	23410
Generated groups (no. of proteins in the largest group)	2,312 (50)	3,108 (19)	1,325 (145)	2,009 (904)
Grouped proteins (% of proteome size)	6,239 (23.4 %)	7,587 (29.7 %)	4,432 (17.9 %)	6,043 (25.8 %)

Tab.5 Evaluation of `grouping` results; comparing the primate and the rodent lineage and the data-sets of NCBI and ENSEMBL
`grouping` parameters: $E \leq 10^{-20}$, Identity ≥ 75 %, Coverage ≥ 80 %

Table 6 presents the results of the `group` output files, generated for the localisation of inner-lineage differences. More proteins could be clustered into groups with the human and mouse proteome respectively. As expected, fewer and smaller groups were generated under these conditions. The located groups should represent gene-families evolved since the last common ancestor of human and chimp in the primate lineage and mouse and rat in the rodent lineage.

	Primates (ENSEMBL)		Rodents (NCBI)	
Organism	**human**	**chimp**	**mouse**	**rat**
Proteome size [no. of sequences]	25578	24376	24728	20080
Generated groups (no. of proteins in the largest group)	761 (10)	316 (8)	681 (45)	502 (24)
Grouped proteins (% of proteome size)	1,621 (6.3 %)	658 (2.7 %)	2,041 (8.3 %)	1,257 (6.2 %)

Tab.6 Evaluation of `grouping` results; comparing human with chimp and mouse with rat
`grouping` parameters: $E \leq 10^{-20}$, human-chimp: Identity ≥ 98 %, Coverage ≥ 98 %,
mouse-rat: Identity ≥ 90 %, Coverage ≥ 95 %

The analysis of the `group` output files clarified that there is a problem with the comparability of the data-sets, derived from the two databases (Tab. 5). Although the proteomes of one organism should be identical in both databases, there are significant differences. These discrepancies are probably a sign of a rather limited knowledge about mammalian proteomes.

Another problem is that the proteome sizes of the organisms vary. This can lead to diverse numbers of generated groups. Indeed more proteins were clustered if the proteome was larger (Tab. 5 and 6).

In the case of the chimpanzee proteome (Tab. 6), less than half of the groups (312) could be generated compared with the human proteome (683 groups). Similar observations were made with the rat and the mouse proteome, noticeably fewer rat proteins, 2035 mouse proteins and only 1250 rat proteins were grouped into gene-families . As these organisms, especially human and chimpanzee are very close relatives, it looks implausible that there is a biological background for these discrepancies. Apparently it is too early to compare the proteomes of human and chimp or rat and mouse at this time, because the knowledge about the chimp and rat proteome (Tab. 1) is still incomplete.

Organism	Identity - Coverage [%]	Database	No. of members	Annotation - Function
human	75 - 80	NCBI	50	No PFAM/PROSITE annotation found
		ENSEMBL	19	Core histone H2A/H2B/H3/H4 (*PFAM*)
	98 - 98	ENSEMBL	13	No PFAM/PROSITE annotation found
chimp	98 - 98	ENSEMBL	10	Core histone H2A/H2B/H3/H4 (*PFAM*)
mouse	75 - 80	NCBI	145	Spin/Ssty Family (*PFAM*)
		ENSEMBL	904	L1 transposable element (*PFAM*)
	90 - 95	NCBI	45	Spin/Ssty Family (*PFAM*)
rat	90 - 95	NCBI	24	Lipocalin signature (*PROSITE*)

Tab.7 Annotations (PFAM/PROSITE) of the largest groups in the `group` output files

Table 7 represents a list of the largest gene-families in the human, mouse, chimp and rat proteome evolved in the specified time interval. The gene-families were located with the `grouping` software. In some cases no PFAM and PROSITE annotation could be found for the groups. This means that until now the biological function of these families is unknown. Furthermore, Table 7 points out that the largest gene-families differ between the NCBI and ENSEMBL data-sets, both in their number of members and in their putative function.

3.2.3.Human and mouse gene-families with identical sequences

The birth of new genes is of interest because it provides raw material for adaptive evolution, with extra copies of genes able to undergo functional divergence in response to positive selection [11].

As it is not possible to compare human gene-families with chimp families at this stage in the genome projects, very recently evolved human families were generated with both data-sets (NCBI and ENSEMBL) for further investigations. These gene-families consist of members with identical amino-acid sequences. In addition, the same analyses were carried out with the mouse proteome, to locate young mouse families.

Organism	Database	Total no. of proteins	No. of groups	No. of proteins in largest Group
Human	NCBI	126	51	14
	ENSEMBL	231	113	3
Mouse	NCBI	211	89	12
	ENSEMBL	1074	268	443

Tab.8 Results of `grouping` with NCBI and ENSEMBL data-sets
Identiy = 100 % and Coverage = 100%

Table 8 shows that there are important differences in the number of total grouped proteins and in the number of groups, between the two human data-sets. The largest group of proteins with identical sequences in the NCBI data-set holds 14 members, in the ENSEMBL data-set just 3. The mouse NCBI-data-set contains nearly twice as much identical proteins than the human NCBI-data-set, but the largest group is similar in size (human: 14; mouse: 12). The mouse data-set obtained from ENSEMBL contains more than five times as many identical proteins as the mouse NCBI data-set.

The annotations/functions of the largest groups in the human data-sets are listed in Table 9 and 10. All generated gene-families, including families with two members, were compared to find identical families between the two data-sets. Eight families were identical and four very similar; up to five amino acids differed between the sequences of these groups. Only one larger group was identical (group no. 7, Tab. 9 and no. 2, Tab. 10) with 3 members and the annotation 'Cystine-knot domain'. The other larger groups (member > 2) of the NCBI data-set were compared with the whole non-redundant ENSEMBL data-set. But no identical or very similar sequence could be found. Conversaly, for the five ENSEMBL groups, at least one identical or very similar sequence was found in the non-redundant NCBI data-set.

No.	Annotation – Function (Human – NCBI)	Proteins	Space in Genome (no. of nucleotides)
1	hTAFII28-like protein conserved region	14	Chr 5: 59,993
2	Ubiquitin carboxyl-terminal hydrolase	7	Chr 4: 39,556
3	PREDICTED: similar to Williams Beuren syndrome	4	Chr 7: 12,562,671
4	No PFAM/PROSITE/NCBI annotation found	3	Chr 1: 12,483
5	Homeobox domain	3	Chr 10: 15,135
6	hTAFII28-like protein conserved region	3	Chr 5: 38,370
7	Cystine-knot domain	3	Chr 19: 25,853
8	Neurotransmitter-gated ion-channel ligand	3	Chr 15: 724,563

Tab.9 Human gene-families with identical amino-acid-sequences (NCBI data-set)
Gene-families with a wide distribution over the chromosome are indicated in yellow;
Annotation: PFAM/PROSITE/NCBI;
Identiy = 100 % and Coverage = 100%, no. of group members > 2;

No.	Annotation – Function (Human – ENSEMBL)	Proteins	Space in Genome (no. of nucleotides)
1	No PFAM/PROSITE annotation found	3	Chr 5 and 8
2	Cystine-knot domain	3	Chr 19: 26,236
3	7 transmembrane receptor (rhodopsin family)	3	Chr 1 and 5
4	Alpha amylase, catalytic domain	3	Chr 1: 102,391
5	TPR Domain	3	Chr 2: 1,141,489

Tab.10 Human gene-families with identical amino-acid-sequences (ENSEMBL data-set)
The only family forming a cluster of nearby genes is indicated in mauve and the two families that are distributed over two chromosomes are indicated in pale green.
Annotation: PFAM/PROSITE
Identiy = 100 % and Coverage = 100%, no. of group members > 2;

In a second step the location of these very young families in the genome was determined and the required space (no. of nucleotides) was calculated, because the Human Genome Sequencing Consortium [11] searched for clusters of nearby homolgous genes as an indication of gene birth.

The sequences of each group, generated with the NCBI data-set (Tab. 9) are located on the same chromosome, respectively. In most of the cases, the sequences of a group were clusters of nearby genes. This was expected because such clusters are indications of recent local gene-dublications. Two of the NCBI-gene-families (No. 3 and 8) are wider distributed over the chromosome. The localisation of the ENSEMBL-groups (Tab. 10) shows that these families are in most of the cases not clusters of nearby genes. Only the sequences of group 2, which is identical to the NCBI group 7, are direct neighbours. Two of the groups are distributed over two different chromosomes.

Based on these results it is difficult to make a statement about the very recently evolved gene-families in the human proteome. In all likelihood there are very young families in the human proteome, but the quality of the current sequence makes it difficult to study this question. According to the Human Genome Sequencing Consortium [11] an increase of gene-duplications in the last 3-4 million years can be detected. But there are several possible explanations for this observation. It may reflect a true increase in the rate of gene duplication in the primate lineage, but on the other hand, these new genes might reflect the transient of duplicated genes and are destined to be culled due to lack of functional benefit.

For the mouse proteome it is even more difficult to get reliable information about recently evolved gene-families. The gene-families obtained from the NCBI data-set are listed with their annotations in Table 11. The largest group of identical sequences in the ENSEMBL data-set has the PFAM annotation 'L1 transposable element', which means that these sequences are repetitive elements that sometimes exhibit a reverse transcriptase acitivity (data not shown).

As long as the two databases (NCBI and ENSEMBL) are so contradictory in their declarations about the type and the number of proteins in the proteomes, it is not possible to perform a reliable analysis of gene birth neither in the human lineage nor in the mouse lineage.

No.	Annotation – Function (Mouse – NCBI)	Proteins
1	Cor1/Xlr/Xmr conserved region	12
2	Zinc finger, C2H2 type	5
3	Spin/Ssty Family	5
4	Glyceraldehyde 3-phosphate dehydrogenase	5
5	Spin/Ssty Family	4
6	Ribosomal protein S8e	3
7	Zinc finger, C2H2 type	3
8	Cor1/Xlr/Xmr conserved region	3
9	Spin/Ssty Family	3
10	Spin/Ssty Family	3
11	Spin/Ssty Family	3
12	Spin/Ssty Family	3
13	Spin/Ssty Family	3
14	No PFAM/PROSITE annotation found	3
15	Cor1/Xlr/Xmr conserved region	3
16	Cor1/Xlr/Xmr conserved region	3

Tab.11 Mouse gene-families with identical amino-acid-sequences (NCBI data-set)
Annotation: PFAM/PROSITE
Identiy = 100 % and Coverage = 100%, no. of group members > 2;

3.3.Comparison of mammalian gene families

3.3.1.Evaluation of the group matching with two different BLAST programs

The major challenge in developing this software was to find a method to match the gene-families of the two organisms. For this purpose the performance of two approaches, both based on BLAST searches was analysed using human and mouse as example organisms (-> 2.5.3). A BLAST search in both directions (human -> mouse & mouse -> human), using the gene-families generated with the `grouping` parameters I ≥ 75 % and C ≥ 80 %, was carried out to test the reliability of the approaches.

Both approaches (`blastall-blastp` and `blastpgp`) detected gene-families without matches to the groups of the other organism. These families are of particular interest as they probably evolved since the split of the compared lineages. Figure 15 shows the number of those groups and the consensus between the two approaches. The approaches found a similar number of groups but the intersection is not 100 % for both organisms. About 10 more unmatched groups were detected in the mouse proteome.

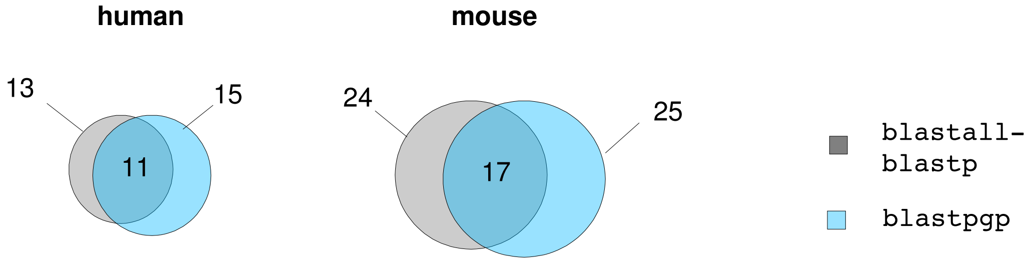


Fig.15 Comparing `blastall-blastp` and `blastpgp` results on the basis of the no. of unmatched groups.
The sets represent the no. of gene-families with no BLAST match to the grouped proteins of the other organism. The intersection represents identical gene-families.
Minimum no. of proteins in group: 10;
`blastall-blastp`-parameters: $E \leq 10^{-20}$, $I \geq 60$ %, $C \geq 80$ %;
`blastpgp`-parameters: $E \leq 10^{-50}$

All BLAST matches between gene-families were clustered (Fig. 5) and the total numbers of proteins in such a cluster were summed up for the two organisms. The clustering step was carried out in both BLAST directions and the results were checked for correlation (Fig. 16). The scheme shows that the clusters received with the `blastall-blastp` approach are identical for both directions, the intersection is 100 %. In contrast to these results, only 30 % of the gene-family clusters could be retrieved with the `blastpgp` approach.

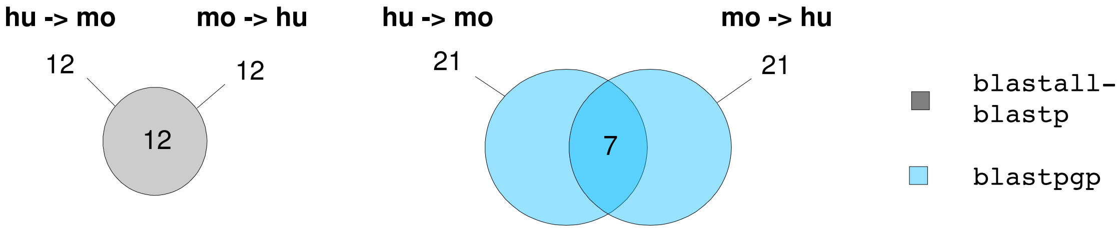


Fig.16 Comparing `blastall-blastp` and `blastpgp` results on the basis of the no. of generated clusters.
The sets represent the no. of gene-family clusters with significant differences in the no. of members. The intersection represents identical gene-family clusters in both BLAST directions
Minimum no. of total grouped proteins of the organism with more proteins in the cluster: 15;
Minimum ratio: larger/smaller no. of total clustered proteins: 1.5;
`blastall-blastp`-parameters: $E \leq 10^{-20}$, $I \geq 60$ %, $C \geq 80$ %;
`blastpgp`-parameters: $E \leq 10^{-50}$

The advantage of the blastpgp approach was the possibility of giving a profile of the gene-family sequences as additional information to the matching step. The poor performance of this approach may have several reasons. First of all, using the expectation value as the sole criterion for identifying homologous proteins is not recommended (-> 2.3.1 parser). Although a profile of the gene-family is given, it might be a problem that only one sequence of a group is chosen for the BLAST search. As described above (-> 1.1) it is possible that only the orhologues match between two organisms. As the sequence was chosen randomly a distant related paralogue might have been chosen for the matching step and as a result no counterpart could be found in the other organism. This leads to the third problem that at least one match must be found in the first round to get the ability of finding more similar sequences, maybe members of other groups, during the next iteration. If in the other direction a more distant paralogue was chosen and no hit was found in the first round the whole cluster would have been lost.

The blastall-blastp matching appears to be a very reliable approach (Fig. 16). As a BLAST search of all against all grouped proteins was carried out the direction was irrelevant and did not cause any problems.

The blastall-blastp approach solved the matching problem much better than the blastpgp approach and was chosen for the final software package.

3.3.2.Comparison of the human and the mouse proteome

The developed software was used to find gene-families in the human and the mouse proteome that expanded since the split of the rodent and the primate lineage and thus differed in size. Gene-families with no match to a group in the other proteome were listed by the software. Clusters of gene-families matching a smaller gene-family in the other organism were listed additionally. Table 12 contains a list of all human gene-families with at least ten members that did not hit a mouse group. The protein sequences of these families were also searched against the whole mouse proteome to demonstrate the evidence of gene duplication in these families. Indeed, in all cases not more than two orthologues were detected in the mouse proteome (last column, Tab. 12).

More unmatched mouse families (25) than human families (13) were generated and the majority of the gene-family clusters contained more mouse proteins (8 of 11, data not shown).

No.	Annotation - source	Group-size	Matches to whole mouse proteome
1	no Pfam/Prosite annotation found	50	0
2	hTAFII28-like protein conserved region-Pfam	25	1
3	Collagen triple helix repeat (20 copies)-Pfam	19	1
4	no Pfam/Prosite annotation found	15	0
5	Sushi/CCP/SCR domain profile-Prosite	14	1
6	Ion transport protein-Pfam	14	0
7	Protein kinase domain profile-Prosite	12	1
8	Ubiquitin carboxyl-terminal hydrolases family 2 profile-Prosite	11	0
9	KRAB-related domain profile-Prosite	10	2
10	Cecropin family signature-Prosite	10	0
11	Zinc finger ZZ-type profile-Prosite	10	0
12	no Pfam/Prosite annotation found	10	1
13	Ribosomal protein S26e signature-Prosite	10	1

Tab.12 Human groups (NCBI) with no BLAST match to a mouse group.
The last column gives the number of BLAST matches to a database of the whole mouse proteome. A gray font is chosen for groups that are no gene-family because the members are splicing variants (Tab. 13).
Parameters: `blastall-blastp` E ≤ 10^{-20}, `parser-group` I ≥ 60 %, C ≥ 80 %;
Minimum no. of proteins in group: 10;

The largest human gene-family (group-size: 50) with no match to the mouse proteome (Tab. 12) could not be annotated. In the `grouping` analysis this family was located as largest family in the human proteome (Tab. 7) that evolved in the last 75 million years. For another two gene-families (Tab. 12, yellow) it was not possible to get information about the function of the members. The proteins in the three groups, indicated in blue are transcription factors, the two groups shown in mauve have an immune function.

Table 13 lists the same groups as Table 12, but gives additional information about the location of these families on the human genome. The table shows that half of the generated unmatched groups (gray font) consist of splicing variants of one to three gene-loci. Splicing variants cover the same gene locus on a chromosome and have a similar starting position. The two largest gene-families (No. 1 and 2) require little space on the chromosome. The other families are wider distributed over one or two chromosomes, but only the members of the last two families (No. 12 and 13) are spread over the whole genome.

No.	**No. of members**	**No. of Prots/ Chromosome**	**Space-size (no. of nucleotides)**	**Comment**
1	50	all/**1**	113,624	
2	25	all/**5**	128,473	
3	19	all/**10**	-	splicing variants of one gene
4	15	(1/**12**) 14/**8**	310480 + 48897 + 18304	total space (8): 7,363,974
5	14	all/**1**	-	splicing variants of one gene
6	14	all/**17**	-	splicing variants of one gene
7	12	all/**1**	-	3 different genes with splicing variants
8	11	7/**4** 4/**8**	Chr4: 39556 Chr8: 1592 + 1592 + 6343	total space (8): 6,874,377
9	10	all/**X**	301,450 + 113,028	9 different genes (one gene with two splicing variants) total space: 10,892,044
10	10	all/**2**	-	splicing variants of one gene
11	10	5/**2;** 5/**18**	-	2 different genes with splicing variants
12	10	2/**1;** 1/**3, 4, 5, 6, 7, 11, 16, 19**	-	
13	10	2/**X**, **8;** 1/**6, 7, 9, 12, 13, 17**	-	

Tab.13 Location of the unmatched human groups (NCBI) on the genome
The colours are inherited from Table 12. A gray font is chosen for groups that are no gene-family because the members are splicing variants.
Parameters: `blastall-blastp` E ≤ 10^{-20}, `parser-group` I ≥ 60 %, C ≥ 80 %;
Minimum no. of proteins in group: 10;

The formation of families distributed over a multiplicity of chromosomes (Tab. 13, No. 12 and 13) is contrary to the theory of the Human Genome Sequencing Consortium [11] which said, only local gene duplications indicate the development of a new gene-family. A more detailed biological study of the function of these families might clarify whether the grouped proteins form a true gene-family or not.

This Consortium [11] analysed gene duplications since the rodent/primate split. They found that duplications are enriched in genes with immune and olfactory function, as well as those likely to be involved in reproductive functions.

With the `comparing` software two human groups (members ≥ 10) with an immune function were generated (Tab. 12, mauve), but these groups are no gene-families as the members are splicing variants of one gene-locus (Tab. 13, mauve).

Human gene families with an olfactory function are not contained in Table 12. A reason for this might be, that our software compares the size of gene-families. Duplication of olfactory receptor genes occurred frequently in both rodent and primate lineages, but overall a larger fraction of mouse genes evolved. Acuity of the olfactory sense is reduced in humans when compared with mouse [4]. Actually seven of the twenty-five unmatched mouse gene-families and four of eight mouse clusters are annotated with 'G-protein coupled receptors family' (data not shown). This superfamily includes among many others, the 'olfactory receptor family' and 'vomeronasal receptor genes', for which the difference in gene-family expansion, according to Dehal et al. [4] is even more clear.

Furthermore Dehal et al. found lineage-specific differences in the coding capacity of gene-families encoding putative transcription factors. Using the `comparing` software, two human gene-families involved in transcription processes were located (Tab. 12, blue; No. 11 is no gene-family -> Tab. 13).

The results of the `comparing` software seem to be reliable, as many observations previously made by other research groups could have been replicated. Our analysis was focused on large gene-family expansions, thus only critical changes were described. Admittedly, it is necessary to work with a data-set containing only the gene-loci, not all transcripts of the loci. Otherwise groups are generated that are only clusters of the transcripts of one gene-locus.

The comparison of two proteomes seems to be feasible with the `comparing` software, but a high quality of the proteome sequence is the basis of a reasonable analysis, especially for studies with very recently splited organisms.

3.4.Software package `ggc`

All programs necessary for the generation and comparison of gene-families were combined into a software package (`ggc` – **g**ene-families: **g**rouping & **c**omparing). Table 14 is a list of all files contained in the package.

Two test-data-sets, two proteomes of *Staphylococcus aureus* are given additionally, to help users getting familiar with the software package. The README file gives some example commands clarifying the usage of `grouping` and `comparing`.

File		Short description
README		Information about functionality and usage of `ggc`
NC_002758.faa		Test data-sets (NCBI): two proteomes of *Staphylococcus aureus*
NC_003923.faa		
grouping		generates gene-families
parser_ncbi	parser_embl	sub programs of major program: `grouping`
group_ncbi	group_embl	
getAllSeq_ncbi	get_allSeq_embl	
getOneSeq_ncbi	get_oneSeq_embl	
parser_prosite		
comparing		compares gene-families of two organisms
parser_group		sub programs of major program: `comparing`
parserN_noMatch	parser_noMatch	
oneSeq		
cluster_group		
cluster2_group		

Tab.14 List of files contained in the software package ggc

4.Summery and Conclusion

During this project a software was developed for grouping and comparing gene-families of two organisms. Paralogous sequences of an organism were grouped together in a gene-family. Afterwards, the sizes of gene-families of two organisms, with orthologous genes were compared to find lineage specific differences. A number of software modules were developed and subsequently merged with public software tools like BLAST or HMMER into two major programs.

The first major program, `grouping` (Fig. 6), combines all the steps that are necessary to generate the groups of paralogous proteins. The grouping of the amino acid sequences can be carried out with fasta files obtained from ENSEMBL or NCBI. For the similarity search the BLAST software was used, to locate all homologous proteins in the proteome. Coverage and identity between the sequences of all matches were calculated and only protein-pairs that fulfilled the homology criteria were passed to the `group` program. At this point the protein-groups were generated using the depth-first-search (DFS) algorithm, which accounts for the transitivity of homology. The last step was searching the PROSITE-database for annotations of the generated gene-families.

The second major program, `comparing` (Fig. 7), integrates all steps necessary to compare the size of gene-families generated with `grouping`. Differences in the size of gene-families of two organisms or gene families with no match to a group of the other organism were located. The matching step was carried out doing a BLAST search with all grouped sequences of both organisms to make sure that no orthologous relationship was missed. A BLAST search in both directions (org1 -> org2 and org2 -> org1) was necessary to confirm the clustering results and to find the unmatched groups of both organisms. Again, coverage and identity were used as criteria for two proteins to be homologous and using the DFS, all matches between the groups (Fig. 5) were clustered together. In a final step all clustered or listed groups without a PROSITE annotation were searched against the PFAM database to get an idea of the group function.

Both software applications were tested and combined, together with two test-data-sets and a file with some example commands, into the software package `ggc`. Both for `grouping` (-> 3.1) and for `comparing` (-> 3.3) the test results show that the performance of the software is reliable.

A major advantage of the `ggc` software is the possibility of generating and comparing gene-families that have diverged in a user-specified time-interval. However, the software is better used for comparisons of not to distantly related lineages, as to relaxed parameters (E-value, identity or coverage) may lead to false classifications (-> 3.1.2). The software may have difficulties locating highly divergent gene-families, but the software was designed to compare closely related lineages like, for example, primates and rodents.

The analyses of the mammalian proteomes was still very difficult, allthough the quality of the proteome sequences of the investigated organisms had been improved a lot in the last years. Problems became obvious on closer examinations of the conflicting data-sets obtained from the NCBI and ENSEMBL database (Tab.1, Tab.5, -> 3.2.3). Furthermore, the necessity of using a data-set containing only the gene-loci, not all transcripts of a locus became apparent (-> 3.3.2).

In summery it was shown that the software performance is satisfactory and `ggc` can be employed for analysing and comparing genomes from different species.

5.List of Abbreviations

AA	Amino acid
BLAST	Basic local alignment search tool
C	Coverage
DFS	Depth first search
EMBL	European Molecular Biology Laboratory
EBI	European Bioinformatics Institute
E-value	Expectation value
HMM	Hidden markov models
I	Identity
NCBI	National Center of Biotechnology Information
No.	Number
Org	Organism
PSI-BLAST	Position-specific iterated BLAST
Seq	Sequence
SRS	Sequence retrieval system

6.Bibliography

[1] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ, Basic local alignment search tool. J Mol Biol 215, 403-410 (1990)

[2] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman, A new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402 (1997)

[3] Bateman A., Coin L., Durbin R., Finn R. D., Hollich V. et al., The Pfam Protein Database. Nucleic Acids Research 32, 138-141 (2004)

[4] Dehal P., Predki P., Olsen A., Stubbs L. et al., Human Chromosome 19 and Related Regions in Mouse Science 293, 104-11 (2001)

[5] Eddy S. R., Profile Hidden Markov Models - review of HMMs Bioinformatics 14, 9 (1998)

[6] Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. R. D. Clayton, F, The minimal gene complement of Mycoplasma genitalium. Science 270, 397-403 (1995)

[7] Gattiker A., Gasteiger E., Bairoch A., ScanProsite: a reference implementation of a PROSITE scanning tool. Applied Bioinformatics 1, 107-108 (2002)

[8] Gu, Z., Cavalcanti, A., Chen, F.-D., Bouman, P. & Li, W.-H., Role of duplicate genes in genetic robustness against null mutations Nature 421, 63-66 (2003)

[9] Gu, Z., Cavalcanti, A., Chen, F.-D., Bouman, P. & Li, W.-H., Extend of gene duplication in the genomes of Drosophila, nematode and yeast Mol. Bio. Evol. 19, 256-262 (2002)

[10] Heun, V., Grundlegende Algorithmen Vieweg 158-161 (2000)

[11] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. Nature 431, 931-945 ()

[12] Li W. H., Molecular evolution. Sinauer (1997)

[13] Mouse Genome Sequencing Consortium., Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-562 (2002)

[14] Rat Genome Sequencing Project Consortium, Genome sequence of the Brown Norway rat yields insights into mam. evol. Nature 428, 493-521 (2004)

[15] Rost, B., Twilight zone of protein sequence alignments. Protein Eng. 12, 85-94 (1999)

[16] Schrago, C.G., Russ, C. A. M., Timing the Origin of New World Monkeys Mol. Biol. Evol. 20, 1620-1625 (2003)

[17] Sigrist C.J.A., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M., Bai, PROSITE: a documented database using patterns and profiles as motif descrip Brief Bioinform. 3, 265-274 (2002)

[18] The Genome International Sequencing Consortrium, Initial sequencing and analysis of the human genome Nature 409, 860-921 (2001)

[19] Wall L., Christiansen T. Schwartz R.L., Programming Perl. O'Reilly, 2nd edition (1996)

7.List of Figures

8.List of Tables

9.Appendix

9.1.List of URLs

EMBL-Sequence retrieval system (SRS)	http://srs.embl-heidelberg.de:8000/srs5/:
NCBI-server	ftp://ftp.ncbi.nih.gov
ENSEMBL-server	ftp://ftp.ensembl.org/pub
PROSITE software and ps_scan.pl	ftp://ftp.expasy.org/databases/prosite/
PFAM and HMMER software	ftp://ftp.genetics.wustl.edu/pub/eddy/
BLAST software	ftp://ftp.ncbi.nih.gov/blast/executables/