IgFamily v0.12.3

Flinders Proteomics Facility

Contents

[Preface 4](#_Toc466980462)

[Story 5](#_Toc466980463)

[- Data: what is it? 5](#_Toc466980464)

[- Data: where does it come from? 5](#_Toc466980465)

[- Data: when is it good? 6](#_Toc466980466)

[- Data: what does it mean? 7](#_Toc466980467)

[- There’s method in models 8](#_Toc466980468)

[Function 15](#_Toc466980469)

[- Cross platform compatability 15](#_Toc466980470)

[- FASTA utility tools 15](#_Toc466980471)

[- Peptide file compatibility 16](#_Toc466980472)

[- msconvert external integration 16](#_Toc466980473)

[- NOVOR v1.1 external integration 16](#_Toc466980474)

[- Local directory functionality 16](#_Toc466980475)

[- Filesystem functionality 17](#_Toc466980476)

[- Runtime user interface 17](#_Toc466980477)

[- FASTA data structuring 18](#_Toc466980478)

[- Peptide data structuring 18](#_Toc466980479)

[- Homology data structuring 18](#_Toc466980480)

[- Protein scoring 18](#_Toc466980481)

[- Probable germlime determination 19](#_Toc466980482)

[- Model conjugation 19](#_Toc466980483)

[- Report generation 19](#_Toc466980484)

[- IgFamily v0.12.3 pseudocode 27](#_Toc466980485)

[Consideration 32](#_Toc466980486)

[- Example: B02+B02a - Sjogren’s syndrome serum with HAGG pulldown 33](#_Toc466980487)

[- Where do peptides belong? 38](#_Toc466980488)

[- From proteins to peptides 42](#_Toc466980489)

[- The problem with parameters 43](#_Toc466980490)

[- A distinct peptide by any other name 47](#_Toc466980491)

[- Distance is relative 47](#_Toc466980492)

[Future 49](#_Toc466980493)

[- Analysis of data generation and reproduction 49](#_Toc466980494)

[- Analysis of germline divergence 49](#_Toc466980495)

[- Blastp internal integration 49](#_Toc466980496)

[- Blastp custom substitution matrix 49](#_Toc466980497)

[- Blastp custom conservation weighting 50](#_Toc466980498)

[- Advanced statistical modellng 50](#_Toc466980499)

[- User interface and interaction 50](#_Toc466980500)

[- Data filesystem 50](#_Toc466980501)

[- Further automation 50](#_Toc466980502)

[- De novo and database process comparison 50](#_Toc466980503)

[- De novo and database process combination 50](#_Toc466980504)

[- Contaminants report and exclusion list generation 50](#_Toc466980505)

[Production 52](#_Toc466980506)

[References 53](#_Toc466980507)

[Workshop 54](#_Toc466980508)

# Preface

Mass spectrometry immunoproteomics has the potential to investigate serum repetoires as an accurate and high-throughput method. Particularly applicable to patients with immunological maladies, immunoglobulins are able to be selectively purified using an antigen or affinity-specific molecule. Typically a sample of immunoglobulins is digested with an enzyme to produce peptides that are suitable for mass spectrometry analysis. Such a sample is able to generate many thousands of spectra that are representative of the peptides present. The entirety of spectra allow insight into the origin proteins of these peptides and the abundance of proteins in the sample. Conventional analysis of detected peptides involves determining those peptides that best represent the possible proteins, often from a database of established proteins, and assigning some metric of the basis of spectral quality, spectral count, supporting peptides, and overall protein sequence coverage.

Although this routine method has proven successful in general proteomics studies, there are complications implicit for immunoproteomics. In particular, the comparative database of immunoglobulins is derived from an ancestral repetition of successful germline alleles known as immunoglobulin gene families. Of these the variable region alleles are responsible for creation of the ~100 amino acid N-terminal and is of importance in diversification of target affinity. There are spatially 80 variable region gene family alleles, each with as many as 20 polymorphic allele variants and potentially many more not yet described. The variable allele has regions of relatively strong germline conservation separated by phylogentically significant diversity. In addition, owing to a RNA polymerase of relatively low mismatch fidelity there are hypervarible regions with a somatically increased nucleotide mutation rate. The generally short-lengthed peptides generated through a immunoproteomics method result in many peptides corresponding to regions of database proteins that are on one case strongly conserved but on the other prone to excessive germline and mutagenic variation. These confounding factors raise concerns of the validity of current mass spectrometry derived immunoproteomic approaches.

The IgFamily program was initially developed to automate the data management and inference workflow of large bodies of mass spectrometry generated data. The concept was to emulate the conventional approach of focusing the inferential effort on distinct peptides representative of the associated proteins. As the program was developed interest grew in the role of supporting peptides and what could be revealed about the holistic complexity of the data.

*IgFamily* is a console based application developed in the C++ programming language.

# Story

## - Data: what is it?

Protein mass spectrometry is capable of producing thousands of spectra from a sample of peptides. Although the spectra are representative of the peptides present, correctly assigning a peptide to a spectrum is a difficult task. The two primary methods of peptide assignment are through database matching and by de novo identification. Database matching relies on an established database for which an in silico enzymatic digest is able to produce peptide candidates that are compared to fragmentation spectra. De novo identification assumes little about the sample and attempts to assign a peptide on the basis of probable amino acid fit. The PEAKS software suite used a combination of de novo and database matching to determine protein likelihood, and is able to produce data for each of these methods independently. The IgFamilyprogram is able to analyse both PEAKS de novo assigned peptides and PEAKS database assigned peptides, as well as peptides assigned by the open-source NOVOR de novo program.

## - Data: where does it come from?

The primary role of the IgFamily program is to determine the proteins present in a sample. Peptides provide evidence of their origin proteins. In a sense, a peptide reveals a snapshot of a region of a protein. Although in general a greater quantity of supporting peptides bolsters the likelihood of a protein, not all peptides have equal discerning power. There are two primary considerations:

- How much of the peptide is present? - What does this peptide belong to?

The first question concerns a complex array of processes. Elution from the chromatographer may result in an otherwise abundant peptide to produce few spectra. Coelution may shroud the quantity of an individual peptide and result in less dedicated mass spectrometer cycle time. Fragmentation efficiency can give a misrepresentative idea of the proportion of peptides present or result in some existing peptides to not be observed altogether. These are a small example of variations that understate the importance of peptide adunbance in determining sample proteins. However, the IgFamily program considers only spectral count as a metric of abundance and encourages the user to keep these complications in mind.

The second question involves the certainty that a peptide can be assigned to a protein given the available evidence. Sequence similarity provides a measure of association to a protein. Considering a potential peptide match requires defining the likelihood of association. The *conditional probability* of a peptide originating from a protein is defined as -

For example, states that the probability that the peptide TAVVYCAR originating from IGHV3-23 is 0.3 or 30%. There exists a distinct conditional probability for each of all peptides to originate from each of all proteins .

For proteins in general the determination of the conditional probability is a routine task. There are often segments of a peptide that are easily distinguishable to a protein even when compared to a large database. Isoform varients have the potential to confound the assignment, although there are likely to be additional supporting peptides to discern between them. However, the proteins of interest for the IgFamilyprogram are immunoglobulins, and the nature of immunoglobulin diversification creates an almost continuum of possible peptide associations. The variable region of an immunoglobulin is partitioned into six defined regions: three framework regions (FR1, FR2, and FR3) separated by three complimentarity-determining regions (CDR1, CDR2, and CDR3). The phylogenetic germline of these regions are known to be more conserved in the FRs, while the CDRs show greater ancestral divergence. In addition, mutation of mature B-cell germlines occurs at an accelerated rate in the CDRs, owing to a RNA polymerase of relatively low mismatch fidelity. As a consequence the spatial location of a peptide has a direct impact on the ability to resolve its origin protein.



Figure 1 - The immunoglobulin gene family ancestry (Frost, et al., 2015). Phylogeny was determined through a maximum-likelihood method

## - Data: when is it good?

An initial step for the analysis of assigned peptides is to filter out data that is not useful or as a worst case spuriously misleading. Consider the PEAKS de novo assigned peptides in Table 1. The PEAKS de novo algorithm assigns a spectral identification along with associated local confidence scores for each proposed residue. Often the terminal ends of a peptide produce poor fragmentation, evident here by local confidence values below 60 for the two leading N-terminal residues. Furthermore, there are assigned peptides with poor C-terminal residues. Two methods are proposed for extracted quality data for the bulk of de novo assignments - Selecting peptides on the basis of average local confidence (ALC) and filtering individual residues by local confidence of a peptide. For a peptide of lengththis is represented by .

Table 1 - Excerpt from peptide\_data report for file WM16\_B02+B02a\_HAGG\_ISOLATED\_RF\_MJ2 showing peptide filtering through a de novo local confidence rolling average method. Here the threshold average value has been set to 85%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| p\_mz | p\_rt | p\_withmod | p\_filtered | p\_denovo\_peptide\_local\_confidence |
| 575.819 | 38.65 | SALVTVSSASTK | LVTVSSASTK | S[45]A[54]L[98]V[99]T[99]V[99]S[98]S[96]A[86]S[89]T[93]K[97] |
| 575.819 | 38.50 | SALVTVSSASTK | LVTVSSASTK | S[44]A[50]L[97]V[99]T[99]V[99]S[98]S[96]A[85]S[88]T[90]K[95] |
| 575.819 | 38.55 | TGLVTVSSASTK | LVTVSSASTK | T[46]G[57]L[98]V[99]T[99]V[99]S[98]S[96]A[86]S[90]T[93]K[97] |
| 575.819 | 38.59 | TGLVTVSSASTK | LVTVSSASTK | T[46]G[52]L[98]V[99]T[99]V[99]S[99]S[97]A[87]S[89]T[93]K[97] |
| 575.822 | 38.40 | SALVTVSSASTK | LVTVSSASTK | S[61]A[50]L[97]V[99]T[99]V[99]S[96]S[89]A[82]S[89]T[94]K[99] |
| 575.822 | 38.45 | SALVTVSSASTK | LVTVSSASTK | S[44]A[51]L[97]V[99]T[99]V[99]S[98]S[96]A[85]S[88]T[93]K[98] |
| 575.822 | 38.55 | SALVTVSSASTK | LVTVSSASTK | S[43]A[51]L[98]V[99]T[99]V[99]S[97]S[92]A[85]S[89]T[94]K[98] |
| 575.822 | 38.59 | TGLVTVSSASTK | LVTVSSASTK | T[69]G[71]L[98]V[99]T[99]V[99]S[98]S[95]A[83]S[90]T[95]K[97] |
| 575.822 | 38.50 | TGLVTVSSASTK | LVTVSSASTK | T[55]G[62]L[97]V[99]T[99]V[99]S[98]S[96]A[86]S[89]T[92]K[95] |

The PEAKS software provides an integrated filtering option through ALC value and is applicable to each of the PEAKS analysis files. Although selection by ALC is simple to implement, it is not robust in extracting the most from the data. In particular, short peptide assignments with very poor local confidences at a few residues will confer a low ALC value, even if there exists a contained subsequence of high confidence. Further still, peptides that have an ALC above the filter threshold may have regions of poor local confidence that can misrepresent the data. The method employed by the *IgFamily* program is to retrieve high local confidence subsequences by use of a rolling average filter, requiring any 3 (less at a terminal) contiguous residues to have an average local confidence greater than a defined value.

The example in Table 1 shows peptides filtered with a rolling average method. Inspecting the peptide assignments of SALVTVSSASTK and TGLVTVSSASTK reveals similar precursor mass-to-charge ratios and retention times, suggesting that these sequences were likely produced by the same originating peptide. The differing factor here is poor N-terminal local confidences. Applying a rolling average method effectively truncates the areas of low local confidence and results in sequences of high quality. It is notable that the produced sequences are identical, showing the reproduction of fragmentation efficiency in these regions of the peptide.

## - Data: what does it mean?

There is still additional knowledge in the data to give evidence to peptide association likelihood. A protein that has accrued substanstial evidence should influence the probability of peptides belonging to that protein. This *prior evidence* is an important consideration when viewing the generated peptides as a sampling of data drawn from the population of peptides - If knowledge exists of the likely peptide distribution *a priori*, this should be represented in the conditional probability. How is all of this achieveable? The answer lies in the creation of a robust statistical model.

## - There’s method in models

The distinctiveness of a peptide can be modelled by relative sequence similarity. In particular, the likelihood of an amino acid substitution is determined largely by physicochemical properties. The *Poisson distribution* (Figure 1) is routinely used to model similarity (Karlin & Altschul, 1990). If sequence similarity is considered as a distance, with query sequences to a more similar subject match being spatially ‘closer’, then the relative distance of a query sequence is approximated by a Poisson model. Strongly distinct peptides by definition are most likely derived from a single or a few proteins and as such the majority of the assignment likelihood is centred on those proteins, represented by a low Poisson rate factor. In contrast, peptides that are not able to be resolved from many proteins have their assignment likelihood shared and a high Poisson rate factor. As a note of technical interest, a high Poisson rate factor approaches a normal distribution *on average*.



Figure 2 - A Poisson distribution with three representative scaling factors. Peptides with greater distinctiveness are likely to have a small scaling factor.

The BLAST (Basic Local Alignment Search Tool) is a primary utility for realising assignment likelihood for a peptide query in comparison to a protein subject database. BLAST relies on a *substitution matrix* as a reference of amino acid substitution. The substitution matrix used for the IgFamily program is the BLOSUM62 matrix. Although routinely used for sequence similarity, it may not necessarily represent the rate of substitution in immunoglobulins. This is examined further in the Considerations section. While useful as a standalone tool the BLAST algorithm is designed as an aid for peptide or protein identification and it is limited in its ability to consider evidence of a sample that may be known or determinable. As a pedagogical example, consider a hypothetical sample where many peptides have already been analysed. Assume that the majority of those peptides are strongly distinct for the gene family IGHV1-69. It is intuitive that further peptides for this sample should be more likely to be assigned to IGHV1-69. Strongly distinct peptides may receive only a smaller relative evidence increase, by nature of their assignment being near-certain initially. However those peptides that are potentially shared among a few peptides will benefit from the evidence contained in the entirety of the sample. Peptides that are heavily shared may need a greater amount of evidence to escape the constraints of a relatively weak assignment. In particular, peptides that are shared in this way may never be confidently assigned as the assignment variance could be greater than the resolving power.



Figure 3 - The BLOSUM62 substitution matrix. Although routinely used for sequence similarity, it may not necessarily represent the rate of substitution in immunoglobulins.

The standard method for adjustment of the peptide assignment Poisson model is with a secondary model that adjusts the model on the basis on prior or determined evidence. Although in general any model has functionality to transform a Poisson, some models are more suitable by their ability to simplify the conjugation. These models are known as *conjugate priors* or *conjugate models*. An appropriate conjugate prior for the Poisson distribution is an exponential distribution.



Figure 4 - An exponential distribution showing three representative scaling factors. The scaling factor describes how strongly prior or determined evidence is considered.

The Poisson-exponential conjugation is sufficient to represent peptide assignment with consideration of evidence determinable from the sample, but how is the evidence itself represented? The canonical model for describing elements that originate from one of many categories is the multinomial distribution. It is clear that any peptide that is observed can only physically result from a single protein. However, the ontological question is how likely such a peptide is a result from one of potentially many proteins. The multinomial distribution allows insight from several equivalent questions, some more intuitive than others -

- If any peptide is sampled, without knowing the peptide sequence, what is the likelihood that it belongs to some protein ?

- If any peptide is sampled, knowing the peptide sequence, what is the likelihood that it belongs to some protein ?

- Without knowing the origin protein , what is the likelihood of sampling a peptide ?

- Knowing the origin protein , what is the likelihood of sampling a peptide ?

Figure 5 - A multinomial distribution for categorical variables. Here the categories are the twenty most likely gene families with non-zero probability for the sample WM16\_B02+B02a\_HAGG\_isolated\_RF\_MJ2. The x-axis shows the gene family and the y-axis shows probability density. Note that the summation of multinomial densities is equal to 1.

The value of multinomial categories is typically the product of a scoring function that assigns a score for each peptide through an appropriate model. The scoring model implemented here is a *general linear model* (GLM). A GLM assigns a score as a linear combination of selected parameters. All combinations of parameters are possible, with some selection of parameters providing a better representation of the reality of the data. Their exists at least one optimal GLM for any observed data set. The GLM used here is the following:

First, determine the homology of the p-th peptide for the g-th protein, as sampled from a Poisson distribution, and repeat for all peptides and proteins . Note that each of all peptides have a strictly positive (although possibly insignificant) homology for each of all proteins -

Determine the transformed homology of the p-th peptide for the g-th and repeat for all peptides and proteins . The transformed homology is The transformed homology represents the weight of relative distance of a peptide to its associatied proteins by considering the homology value, mismatch count, and incomplete alignment count.

Determine the conjugated transformed homology of the p-th peptide for the g-th protein and repeat for all peptides and proteins . The conjugated transformed homology is the product of the transformed homology with the conjugation value . The conjugation value is the value of the g-th protein in proportion to the maximum scoring protein, raised to the power of a non-negative scalar . The eagle-eyed reader will notice that is dependent on which is itself dependent on through . As described shortly, initially , such that , with increasing as the algorithm converges towards a clustering of gene families -

Determine the transformed homology density of the p-th peptide for the g-th protein and repeat for all peptides and proteins . The transformed homology density is the value of the transformed homology of the p-th peptide for the g-th protein in comparison to the sum of all transformed homologies for the p-th peptide for all proteins -

Determine the conjugated homology density of the p-th peptide for the g-th protein and repeat for all peptides and proteins . The conjugated homology density is the value of the conjugated homology of the p-th peptide for the g-th protein in comparison to the sum of all conjugated homologies for the p-th peptide for all proteins -

Determine the peptide score of the p-th peptide for the g-th protein and repeat for all peptides and proteins . The peptide score is a product of the conjugated homology density and the transformed homology density raised to the power of a non-negative scalar . The product represents the Poisson-exponential conjugation, with functionalising the Poisson and functionalising the exponential. The scalar provides weighting to the prior distribution of the Poisson and represents the naivity of the model. A greater will result in values of that do not diverge greatly from the initially assigned transformed homology density . In constrast, as evidence towards the conjugated homology density increases, most notably with a greater sample size (a larger count of observed peptides), will approach zero and the peptide score will be dominated by -

Determine the multinomial value for the g-th protein and repeat for all proteins . The multinomial value is the summation of all the peptide scoresfor all peptides . The multinomial value represents the evidence for the protein as inferred by the model.

Finally, determine the multinomial density for the g-th protein and repeat for all proteins . The multinomial density is the value of the multinomial value for the g-th protein in comparison to the sum of all multinomial values -

The multinomial density is an important measure of inference. In particular, it reflects the question “If any single gene family protein were observed from sample of gene family proteins, what is the likelihood that it is gene family ?”. This value also determines the proportion of gene families present in the sample. Note that the sum of multinomial densities is equal to 1 - This is typical of a full probability distribution (the combination of all possible outcomes in the model is representive of every possible outcome).

With the framework of a model established, the important question of parameter selection is apparent. Without distinct outcome variables there is only the inferable evidence to deduce parameters that are intuitive while also robust. The concept utilised by the IgFamily program is the idea of *maximum entropy*. This can be observed by simulating the model with a variety values for the parameters and relating them to the overall distribution of such models. As a example, suppose a sample was abundant in IGHV1-69 and IGHV3-7 proteins. With a variety of input parameters..

It makes sense to view regression towards a defined number of sample gene families. The reasoning follows from the belief that the majority of the peptides would originate from a small number of gene family proteins (however, the user may select any number of gene families to regress towards or choose not to at all). This is reflected in the strength of peptide association. Peptides strongly distinct for a gene family should only increase in absolute conditional probability density slightly, by virtue of the difference between and absolute certainty being small. The most shared peptides would resolve to the most likely gene family assignments. Here the prior belief of total present gene families comes into force, the model is constrained by the distribution of assignments to only those few that make sense in the model - a peptide shouldn’t have a likelihood of originating from a great number of gene families if only a few are believed to be present in the sample. The model is itself however subject to a degree of scepticism which is represented by the value , a value known as the *prior belief* or *prior distribution* of the model.

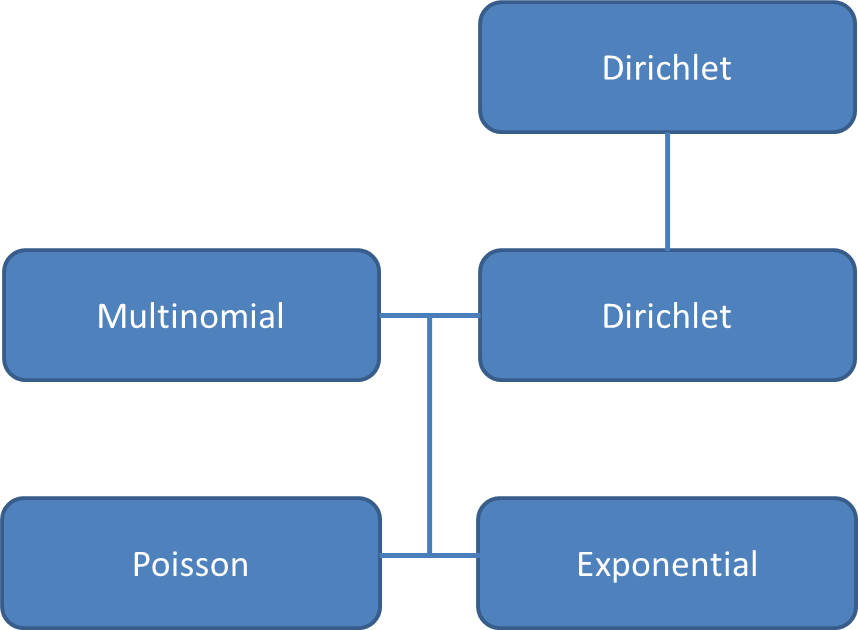
****

Figure 6 – General model form.

The use of belief systems as epistemological concepts to define statistical models is typical of *Bayesian inference*. It may be interpreted as having an arbitrary nature and in a sense this is true. But like with any view of statistical models the choice of model components (likelihood distributions) and parameters are themselves only a utility to make sense of the data. Just as there are some parameters that fit the data better than others, so too are there models that are more representative of the system.

Perhaps a view of such models is mistakenly understood to be a grim herald. It is for any necessarily subjective model that there exists at least one objectively best representative set of parameters to fit the data, provided such parameters can be found. All hope is not lost however, the eminent George Box captured in his timeless aphorism - "The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful".

# Function

The IgFamily program is a console based application developed using the C++ programming language. It operates by reading export files produced by the PEAKS database, PEAKS de novo, or NOVOR de novo programs along with a FASTA file containing a protein database. These upstream programs create a peptide assignment for each of the spectra contained in a mass spectrometry data file. Through this, the IgFamily program creates a data structure of peptide assignments and a reference data structure of proteins and determines the association of these peptides to the proteins by the model described in the Story section. The initial sequence similarity is determined by the NCBI blastp executable which is externally integrated into the IgFamily program.

Functionality is modifiable by the user through a simple console-based interface. The user is able to select which FASTA file(s) to use, create a custom FASTA file, select which peptide assignment file to input, and include the msconvert file conversion and NOVOR de novo programs into the runtime workflow. The resulting output file is designed around two HTML files that display the most probable gene families and associated peptides. A description of key features follows and concludes with the pseudocode of the IgFamily program.

## - Cross platform compatability

The default version of the IgFamily requires a 64-bit Windows operating system and may be compiled for 32-bit Windows, 32-bit and 64-bit Mac OS X, and 32-bit and 64-bit Linux distributions.

## - FASTA utility tools

A subroutine is able to parse FASTA files into runtime. The input format is standardised as:

>[ACCESSION]|[NAME]|[TYPE]|[SPECIES]|

A FASTA creation utility provides functionality for creating custom FASTA files with any combination of fields. This utlility can be used to create the required FASTA format and may be used to include parsing rules for any input format. Accession field is an identifier field and is not required for runtime. Name, type, and species fields are used in data structure creation, association, and function.

## - Peptide file compatibility

Available software is able to determine a peptide assignment for generated spectra. The IgFamily program is able to parse export files created by these programs for downstream analysis. The currently supported software is PEAKS v8.0 DE NOVO de novo peptides .csv export, PEAKS v8.0 SPIDER protein-peptide .csv export, and NOVOR v1.1 de novo peptides .csv export.

From the PEAKS v8.0 DE NOVO de novo peptides export file IgFamily parses scan number, peptide accession (identifier value), peptide with modification, peptide mass-to-charge ratio, peptide retention time, peptide theoretical mass, peptide amino acid local confidence score, and source file name. Note that PEAKS v8.0 DE NOVO assigns individual peptide accessions for replicate peptide assignments.

From the PEAKS v8.0 SPIDER protein peptide file export file IgFamily parses scan number, peptide accession (identifier value), peptide with modification, peptide mass-to-charge ratio, peptide retention time, peptide theoretical mass, peptide spectral count, peptide assignment confidence value IgP, and source file name.

From the NOVOR v1.1 de novo peptides file export file IgFamily parses scan number, peptide accession (identifier value), peptide with modification, peptide mass-to-charge ratio, peptide retention time, peptide theoretical mass, peptide amino acid local confidence score, and source file name. Note that here NOVOR v1.1 assigns individual peptide accessions for replicate peptide assignments.

## - msconvert external integration

The msconvert mass spectrometry data file convertor is able to be called through a user defined interface option to convert AB Sciex .wiff and .wiff.scan file types into the Mascot Generic Format .mgf file type. Various command line options may be selected with peak-picking as the default option. The generated file is created in the same folder as the input file.

## - NOVOR v1.1 external integration

The NOVOR v1.1 de novo command line applicationis able to be called through a user defined interface option to generate NOVOR v1.1 de novo peptide assignment files in the .csv file type. Various command line options may be selected. The generated file is created in the same folder as the input file.

## - Local directory functionality

The IgFamily program is able to be run in local directory or filesystem directory file mode. In local directory file mode the user places the required files in the IgFamilyroot directory and executes the program. The input file and output files are moved to a folder created in a subdirectory with the name of the input file sample. The local file mode is able to be executed from any filesystem location.

## - Filesystem functionality

The IgFamily program is able to be run in local directory or filesystem directory file mode. In filesystem mode the program accessions a dedicated file association structure to retrieve and export files. The filesystem is currently defined on the *FATELVIS* network assisted storage device. Initially the user is required to accession an input file (Figure 6), however there is proposed functionality for dynamic file management with the ISO/IEC TS 18822:2015filesystem library. A useful feature of the filesystem is the addition of data in an accesion file not also included in the input files. In the example in Figure 6 the patient status is primary Sjogrens Syndrome (pSS) - With a large scale study this would allow factor analysis between files with different covariates: patient status, progression of disease at sampling, treatment, mass spectrometry settings, and any other variables of interest.



Figure 7 - An example of a filesystem accession file. Currently the user is required to create an accession file to operate the IgFamily in filesystem mode.

## - Runtime user interface

At initialisation of the IgFamily program the user is greeted with an interactive menu. The user is able to access FASTA file utilities, msconvertcommand line tools, NOVOR v1.1 command line tools, and select program parameters. Program parameters include local or filesystem file modes, FASTA file selection, and peptide assignment selection.



Figure 8 - Through the interface the user can select a FASTA file.

****

Figure 9 - Through the interface the user can select a peptide assignment method.

## - FASTA data structuring

Following user selection of runtime parameters, the IgFamily program will parse the FASTA file(s). The [NAME], [TYPE], and [SPECIES] fields have runtime functionality, although the [ACCESSION] field is retained for FASTA utility functions. An additional data type is created to define those FASTA accessions that are immunoglobulin. An example excerpt is shown in Table 2.

## - Peptide data structuring

Following user selection of runtime parameters, the IgFamily program will parse the data .csv file(s). From this data, the raw peptide creates two additional data types for the peptide with modifications removed and the peptide truncated based on associated de novo local confidence scores. Truncation requires a moving average of 85% for amino acid local confidence and a minimum peptide length of 5. In the event that a peptide is cleaved at a midpoint such that two peptides of length >= 5 are produced and both are assigned to unique data structures. An example excerpt is shown in Table 3.

## - Homology data structuring

Following runtime data structuring of FASTA and peptide data, a blastp reference database is created from the current FASTA data, and an input peptide query list is created from the current peptide data. The input peptide query list is measured for similarity against the blast reference database and a results file is generated as an output file. Blastp is programmed to allow up to 200 matches for each query, and a generous threshold for alignment acceptance. The blastp output file is parsed into runtime and assigned to a homology data structure and peptide queries are associated to their peptide data structure and protein data structure counterparts. With the peptides associated they are able to scored through the model proposed in the Story section. An example excerpt is shown in Table 4.

## - Protein scoring

With the homology data structure created, scored peptides determine the protein scoring as proposed in the Story section. An example excerpt is shown in Table 5.

## - Probable germlime determination

With the homology data assigned and proteins scored the IgFamily program determines the most likely germline allele usage and uses this to a refined blastp database. This is followed with a second iteration of homology scoring and data structuring.

## - Model conjugation

To consider the overall evidence contained in the sample the IgFamily program transforms the peptide homology values with a second conjugate function as proposed in Story section. Adjusted homology values modify peptide scores and consequent protein scores. The conjugation process is iterated until the clustering condition is achieved.

## - Report generation

As a result of its analysis the IgFamily program produces two primary HTML reports: a summary report showing the most likely present gene families with consensus results and an expanded report that additionally includes associated peptides. The gene families are ranked by the multinomial value while the peptides are ranked by the peptide score . In the following pages the summary report is shown along with the expanded report for the top three gene families. Both reports display the consensus information. Note that all proteins in the database are considered while only the gene family proteins are shown. The HTML report includes colour-coding for fast readability (Table 7).

Table 2 - Example excerpt of protein\_data report file.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| key | protein\_name | protein\_type | protein\_species | protein\_protein |
| 0 | IGHV1-18\*01 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGQGLEWMGWISAYNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTAVYYCAR |
| 1 | IGHV1-18\*02 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGQGLEWMGWISAYNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTA |
| 2 | IGHV1-18\*03 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGQGLEWMGWISAYNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDMAVYYCAR |
| 3 | IGHV1-18\*04 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWVRQAPGQGLEWMGWISAYNGNTNYAQKLQGRVTMTTDTSTSTAYMELRSLRSDDTAVYYCAR |
| 4 | IGHV1-2\*01 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLEWMGRINPNSGGTNYAQKFQGRVTSTRDTSISTAYMELSRLRSDDTVVYYCAR |
| 5 | IGHV1-2\*02 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLEWMGWINPNSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTAVYYCAR |
| 6 | IGHV1-2\*03 | IGV | Homo\_sapiens | QVQLVQSGAEVKKLGASVKVSCKASGYTFTGYYMHWVXQAPGQGLEWMGWINPNSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTAVYYCAR |
| 7 | IGHV1-2\*04 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLEWMGWINPNSGGTNYAQKFQGWVTMTRDTSISTAYMELSRLRSDDTAVYYCAR |
| 8 | IGHV1-2\*05 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKASGYTFTGYYMHWVRQAPGQGLEWMGRINPNSGGTNYAQKFQGRVTMTRDTSISTAYMELSRLRSDDTVVYYCAR |
| 9 | IGHV1-24\*01 | IGV | Homo\_sapiens | QVQLVQSGAEVKKPGASVKVSCKVSGYTLTELSMHWVRQAPGKGLEWMGGFDPEDGETIYAQKFQGRVTMTEDTSTDTAYMELSSLRSEDTAVYYCAT |

Table 3 - Example excerpt of peptide\_data output file.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| key | scan\_ID | peptide\_mz | peptide\_z | peptide\_rt | peptide\_m | peptide\_withmod | peptide\_withoutmod | peptide\_filtered |
| 1036 | 19205 | 800.905 | 2 | 41.85 | 1599.81 | A(+27.99)APSGVTTDKVQAEAK | AAPSGVTTDKVQAEAK | SGVTTDKVQAEAK |
| 1675 | 19035 | 800.906 | 2 | 41.33 | 1599.81 | A(+27.99)APSGVTTDKVQAEAK | AAPSGVTTDKVQAEAK | VTTDK |
| 3559 | 27064 | 696.285 | 3 | 38.96 | 2085.85 | A(+27.99)CSVSCGQ(+.98)LCDLLECKDDR | ACSVSCGQLCDLLECKDDR | QLCDLL |
| 1025 | 17823 | 1068.51 | 2 | 36.06 | 2135.00 | A(+27.99)DNALKSGNSKESVTEQDSK | ADNALKSGNSKESVTEQDSK | ALKSGNSKESVTEQ |
| 3420 | 26503 | 719.986 | 3 | 36.31 | 2156.95 | A(+27.99)DNALQSYMNEVSTEQTTK | ADNALQSYMNEVSTEQTTK | ADNALQ |
| 1296 | 16439 | 447.714 | 2 | 30.05 | 893.413 | A(+27.99)EGPTTYK | AEGPTTYK | GPTTYK |
| 1523 | 18849 | 659.798 | 2 | 40.63 | 1317.58 | A(+27.99)ESTAVCLEDPK(+27.99) | AESTAVCLEDPK | AESTAV |
| 3234 | 27467 | 659.790 | 2 | 40.73 | 1317.58 | A(+27.99)ESTAVLCEDPK(+27.99) | AESTAVLCEDPK | ESTAV |
| 911 | 17579 | 809.418 | 2 | 34.83 | 1616.84 | A(+27.99)KGSGVTTDKVQAEAK | AKGSGVTTDKVQAEAK | GSGVTTDKVQAEAK |
| 1104 | 17888 | 809.417 | 2 | 36.28 | 1616.84 | A(+27.99)KGSGVTTDKVQAEAK | AKGSGVTTDKVQAEAK | GSGVTTDKVQAEAK |

|  |  |
| --- | --- |
| denovo\_peptide | denovo\_peptide\_filtered |
| A[39]A[49]P[37]S[94]G[95]V[98]T[99]T[98]D[98]K[95]V[94]Q[92]A[96]E[98]A[93]K[97] | S[94]G[95]V[98]T[99]T[98]D[98]K[95]V[94]Q[92]A[96]E[98]A[93]K[97] |
| A[27]A[31]P[18]S[44]G[75]V[88]T[96]T[89]D[90]K[86]V[81]Q[76]A[82]E[92]A[80]K[91] | V[88]T[96]T[89]D[90]K[86] |
| A[71]C[28]S[26]V[60]S[56]C[20]G[11]Q[87]L[93]C[89]D[93]L[94]L[93]E[86]C[11]K[10]D[56]D[79]R[93] | Q[87]L[93]C[89]D[93]L[94]L[93] |
| A[46]D[61]N[76]A[87]L[95]K[97]S[98]G[91]N[93]S[95]K[93]E[97]S[89]V[91]T[92]E[95]Q[79]D[93]S[65]K[86] | A[87]L[95]K[97]S[98]G[91]N[93]S[95]K[93]E[97]S[89]V[91]T[92]E[95]Q[79] |
| A[86]D[95]N[95]A[96]L[98]Q[95]S[87]Y[31]M[11]N[87]E[96]V[81]S[72]T[82]E[87]Q[65]T[18]T[24]K[62] | A[86]D[95]N[95]A[96]L[98]Q[95] |
| A[32]E[57]G[92]P[90]T[94]T[94]Y[92]K[99] | G[92]P[90]T[94]T[94]Y[92]K[99] |
| A[86]E[93]S[85]T[95]A[95]V[93]C[76]L[79]E[86]D[69]P[24]K[29] | A[86]E[93]S[85]T[95]A[95]V[93] |
| A[82]E[92]S[85]T[96]A[97]V[93]L[66]C[43]E[82]D[85]P[51]K[57] | E[92]S[85]T[96]A[97]V[93] |
| A[35]K[43]G[94]S[98]G[94]V[97]T[99]T[98]D[98]K[95]V[95]Q[95]A[92]E[96]A[92]K[97] | G[94]S[98]G[94]V[97]T[99]T[98]D[98]K[95]V[95]Q[95]A[92]E[96]A[92]K[97] |
| A[37]K[41]G[90]S[97]G[85]V[93]T[98]T[91]D[93]K[89]V[95]Q[84]A[83]E[91]A[87]K[92] | G[90]S[97]G[85]V[93]T[98]T[91]D[93]K[89]V[95]Q[84]A[83]E[91]A[87]K[92] |

Table 4 - Example excerpt of homology\_data report file.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| key\_query | query | subject | key\_subject\_accession | subject\_accession | mismatch\_count | alignment\_coverage\_delta | alignment\_coverage |
| 17 | YYVDSVK | YYVDSVK | 184 | IGHV3-7 | 0 | 0 | 100 |
| 17 | YYVDSVK | YYVDSVK | 160 | IGHV3-52 | 0 | 0 | 100 |
| 17 | YYVDSVK | HYVDSVK | 82 | IGHV3-16 | 1 | 0 | 100 |
| 17 | YYVDSVK | YYADSVK | 165 | IGHV3-53 | 1 | 0 | 100 |
| 17 | YYVDSVK | YYADSVK | 178 | IGHV3-66 | 1 | 0 | 100 |
| 17 | YYVDSVK | YYADSVK | 96 | IGHV3-23 | 1 | 0 | 100 |
| 17 | YYVDSVK | YYADSVK | 147 | IGHV3-43 | 1 | 0 | 100 |
| 17 | YYVDSVK | YYADSVK | 131 | IGHV3-30-5 | 1 | 0 | 100 |
| 17 | YYVDSVK | YYADSVK | 105 | IGHV3-30 | 1 | 0 | 100 |
| 17 | YYVDSVK | YYADSVK | 126 | IGHV3-30-3 | 1 | 0 | 100 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| homology | homology\_transformed | homology\_transformed\_conjugated | homology\_density | homology\_density\_conjugated | score |
| 3400 | 2.29E+12 | 6.76E+11 | 0.47473 | 0.99277 | 26.2091 |
| 3400 | 2.29E+12 | 8.66E-67 | 0.47473 | 1.27E-78 | 3.36E-77 |
| 3200 | 2.74E+10 | 7.79E-130 | 0.00567 | 1.14E-141 | 2.88E-140 |
| 3000 | 2.19E+10 | 4.95E-07 | 0.00453 | 7.27E-19 | 1.74E-17 |
| 3000 | 2.19E+10 | 6.07E-08 | 0.00453 | 8.92E-20 | 2.14E-18 |
| 3000 | 2.19E+10 | 0.000298 | 0.00453 | 4.37E-16 | 1.05E-14 |
| 3000 | 2.19E+10 | 3.84E+08 | 0.00453 | 0.00057 | 0.01355 |
| 3000 | 2.19E+10 | 1.09E+07 | 0.00453 | 1.60E-05 | 0.00038 |
| 3000 | 2.19E+10 | 1.09E+07 | 0.00453 | 1.60E-05 | 0.00038 |
| 3000 | 2.19E+10 | 165496 | 0.00453 | 2.43E-07 | 5.83E-06 |

Table 5 - Example excerpt of protein\_analysis report file.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| key | protein\_name | protein\_score | proteinconstruct\_sequencecoverage | protein\_type | homology\_query | homology\_subject | denovo\_replicate\_count | peptide\_score\_density |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | LVQSGAEVK | LVQSGAEVK | 23 | 0.22022 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | FGTANYAQK | FGTANYAQK | 6 | 0.10390 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | YALSWVR | YAISWVR | 7 | 0.09569 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | YAQNFQGR | YAQKFQGR | 8 | 0.08158 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | VVQSGAEVK | LVQSGAEVK | 7 | 0.05755 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | FATANYAQK | FGTANYAQK | 4 | 0.05347 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | EDTAVY | EDTAVY | 10 | 0.04899 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | FGTANYAQR | FGTANYAQK | 3 | 0.04466 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | QEDTAVY | EDTAVY | 11 | 0.04280 |
| 85 | IGHV1-69 | 329.172 | 51.02 | IGV | QLVQSGAEVK | QLVQSGAEVK | 3 | 0.03137 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| mismatch\_count | alignment\_coverage\_delta | homology | homology\_transformed | homology\_transformed\_conjugated | homology\_density | homology\_density\_conjugated |
| 0 | 0 | 4100 | 4.41E+12 | 1.43E+12 | 0.09472 | 0.61802 |
| 0 | 0 | 4700 | 7.12E+12 | 2.31E+12 | 0.99999 | 0.99999 |
| 0 | 0 | 3500 | 2.54E+12 | 8.19E+11 | 0.97528 | 0.99998 |
| 1 | 0 | 3500 | 3.75E+10 | 1.19E+10 | 0.16743 | 0.74592 |
| 1 | 0 | 3700 | 4.56E+10 | 1.44E+10 | 0.08751 | 0.57576 |
| 1 | 0 | 3400 | 3.39E+10 | 1.07E+10 | 0.99999 | 0.99999 |
| 0 | 0 | 3000 | 1.48E+12 | 4.76E+11 | 0.03638 | 0.40311 |
| 1 | 0 | 3900 | 5.48E+10 | 1.74E+10 | 0.99999 | 0.99999 |
| 1 | 1 | 2800 | 1.52E+09 | 4.72E+08 | 0.02836 | 0.33703 |
| 0 | 0 | 4600 | 6.60E+12 | 2.14E+12 | 0.09436 | 0.61473 |

Table 6 - Information shown in HTML report.

|  |  |
| --- | --- |
| Term | Definition |
| Protein | The protein |
| Score | The value of the multinomial element |
| Density | The density of the multinomial element |

|  |  |
| --- | --- |
| Term | Definition |
| Conjugated density | The value of |
| Density | The value of |
| Conjugated homology | The value of |
| Homology | The value of |
| Local confidence | The value of |

Table 7 - The HTML report includes colour-coding for fast readability.

|  |  |  |
| --- | --- | --- |
| Term | Colour | Range |
| Conjugated density | BLUE  GREEN  ORANGE  RED |  |
| Density | BLUE  GREEN  ORANGE  RED |  |
| Conjugated homology | BLUE  GREEN  ORANGE  RED |  |
| Homology | BLUE  GREEN  ORANGE  RED |  |
| Local confidence | BLUE  GREEN  ORANGE  RED |  |



Figure 10 - Excerpt of summary HTML report for file JW16\_M03+M03a\_HAGG\_RF\_KM\_75kd. The top 3 gene families are shown.



Figure 11 - Excerpt of expanded HTML report for file JW16\_M03+M03a\_HAGG\_RF\_KM\_75kd. The top 30 associated peptides are shown.

## - IgFamily v0.12.3 pseudocode

The pseudocode for the IgFamily program follows. The complete source code can be found at -

https://github.com/Lukah0173/IgFamily

1. Initialise default settings and prompt user to confirm default settings or select custom settings.

2. If selected, perform msconvert file conversion:

2a. If (perform\_wiff\_fileconversion) perform wiff\_fileconversion().

3. If selected, perform NOVOR de novo peptide assignment:

3a. If (perform\_novor\_denovo) perform novor\_denovo().

4. Parse FASTA files into raw data structures:

4a. For (selected\_FASTA\_file) perform parse\_FASTA();

5. Parse data files into raw data structures:

5a. If (peptide\_assignment\_method == PEAKS\_database) perform parse\_PEAKS\_database\_peptides().

5b. If (peptide\_assignment\_method == PEAKS\_denovo) perform parse\_PEAKS\_denovo\_peptides().

5c. If (peptide\_assignment\_method == NOVOR\_denovo) perform parse\_NOVOR\_denovo\_peptides().

6. Assign raw data structures to designed data structures:

6a. Perform create\_v\_peptide\_data().

6b. Perform create\_v\_peptide\_analysis().

6c. Perform create\_v\_protein\_data().

7. From assigned data structures create blastp input file and blastp database:

7a. Perform create\_blastinput().

7b. Perform create\_blastp\_database().

8. Create a system process and direct blastp.exe to created input file and database:

8a. Perform systemcall\_blastp().

9. Parse blastp output file to raw data structures:

9a Perform parse\_homology\_data().

10. Transform homology data and associate homology data to peptide and protein data structures:

10a. Perform transform\_homology\_data().

10b. Perform associate\_homology\_data\_to\_peptide\_data().

10c. Perform associate\_homology\_data\_to\_protein\_data().

11. Through homology data association to peptide and protein data, determine homology\_density and score:

11a. Perform determine\_homology\_data\_parameters().

12. Create protein\_analysis data structures:

12a. Perform create\_v\_protein\_analysis().

13. Determine protein\_analysis parameters:

13a. Perform determine\_protein\_score\_density().

13b. Perform determine\_sequence\_coverage().

13c. Perform sort\_v\_protein\_analysis().

14. Determine most likely germline allele representation and create new blastp input:

14a. Perform select\_protein\_analysis\_by\_score().

14b. Perform create\_blastp\_database\_refined().

15. Create a system process and direct blastp.exe to created input file and database:

15a. Perform systemcall\_blastp().

16. Parse blastp output file to raw data structures:

16a Perform parse\_homology\_data().

17. Transform homology data and associate homology data to peptide and protein data structures:

17a. Perform transform\_homology\_data().

17b. Perform associate\_homology\_data\_to\_peptide\_data().

17c. Perform associate\_homology\_data\_to\_protein\_data().

18. Align query data to subject data:

18. create\_blastp\_query\_alignment().

19. Through homology data association to peptide and protein data, determine homology\_density and score:

19a. Perform determine\_homology\_data\_parameters().

20. Create protein\_analysis data structure:

20a. Perform create\_v\_protein\_analysis().

21. Conjugate homology\_data and protein\_analysis score through iterative process, until cluster condition is achieved:

21a. While (count\_ClusterProportion() > select\_nGeneFamilies) Perform conjugate\_homology ().

22. Determine protein\_analysis parameters:

22a. Perform determine\_protein\_score\_density().

22b. Perform determine\_sequence\_coverage().

22c. Perform sort\_v\_protein\_analysis().

23. Create consensus protein construct for \_ProteinAnalysis data:

23a. Perform create\_ProteinConstruct().

24. Create multinomial data frame:

24a. Perform create\_MultinomialData().

25. Create report and output data:

25a. Perform fout\_v\_PeptideData().

25b. Perform fout\_v\_ProteinData().

25c. Perform fout\_v\_PeptideAnalysis().

25d. Perform fout\_v\_ProteinAnalysis().

25e. Perform fout\_v\_HomologyData().

25f. Perform fout\_Multinomial().

25g. Perform fout\_MultinomialElement().

25h. Perform fout\_MultinomialElementNoMatch().

25i. Perform fout\_MultinomialContaminantsReport().

25j. Perform fout\_MultinomialContaminantsList().

25k. Perform fout\_MultinomialProteinScore().

25l. Perform fout\_MultinomialProteinDensity().

25m. Perform fout\_ProteinPseudoabundanceScore().

25n. Perform fout\_HTMLReport().

25o. Perform fout\_Filesystem().





# Consideration

To clarify the use of the IgFamily program an example is always necessary. The example here is generously provided by collaboration. The functionality of the model is described with detail in the **Story** and **Function**. As a reference, terms of interest are -

Table 8 - Terms associated with the IgFamily program.

|  |  |
| --- | --- |
| Term | Definition |
|  | The peptide of all peptides . |
|  | The protein of all peptides . |
|  | The number of observed spectra assigned as the peptide . |
|  | The protein sequence aligned to by peptide . |
|  | The sequence mismatch count of the p-th peptide for the g-th protein. Each amino acid mismatch of the peptide sequence to the database sequence increases by a value of 1. The value of along with has an important role in determining the transformed homology density and represents the rate of amino acid substition. |
|  | The alignment coverage mismatch count of the p-th peptide for the g-th protein. Occasionally, homology matching will match a subsequence of a peptide sequence to a database match if it generates a higher homology score. This is often a result of a mismatch at the terminus of a peptide sequence to a database sequence. Each amino acid not included in the homology match increases the value of by 1. The value of along with has an important role in determining the transformed homology density and represents the rate of amino acid substition. |
|  | The homology of the p-th peptide for the g-th protein. The initial homology value of the peptide sequence to the database sequence. This value is generated through a consensus view of amino acid substitution represented by a substitution matrix. The substitution matrix used for IgFamily is the BLOSUM62 matrix, routinely used for amino acid sequence similarity. However, it is generated through conserved transmembrane protein sequences and may not be suitable for immunoglobulin substitution rates. |
|  | The transformed homology of the p-th peptide for the g-th protein. The transformed homology is the value of the modified to consider the impact of and as well as an additional scaling factor representing the overall rate of amino acid substitution. |
|  | The conjugated homology of the p-th peptide for the g-th protein. |
|  | The transformed homology density of the p-th peptide for the g-th protein. |
|  | The conjugated homology density of the p-th peptide for the g-th protein. |
|  | The score of the p-th peptide for the g-th protein. |

## - Example: B02+B02a - Sjogren’s syndrome serum with HAGG pulldown

The patient has primary Sjogren’s syndrome and several sera samples have been collected. The sample B02+B02a is a HAGG pulldown technical duplicate of a serum sample collected before treatment with.. The gene family summary report is shown in Figure 11. The expanded reports for the top three ranking gene families are shown in Figure 12, Figure 13, and Figure 14. The three highest scoring gene families are IGHV1-69, IGHV3-7, and IGHV3-69-1, with multinomial densities of 0.225, 0.215, and 0.142 respectively.

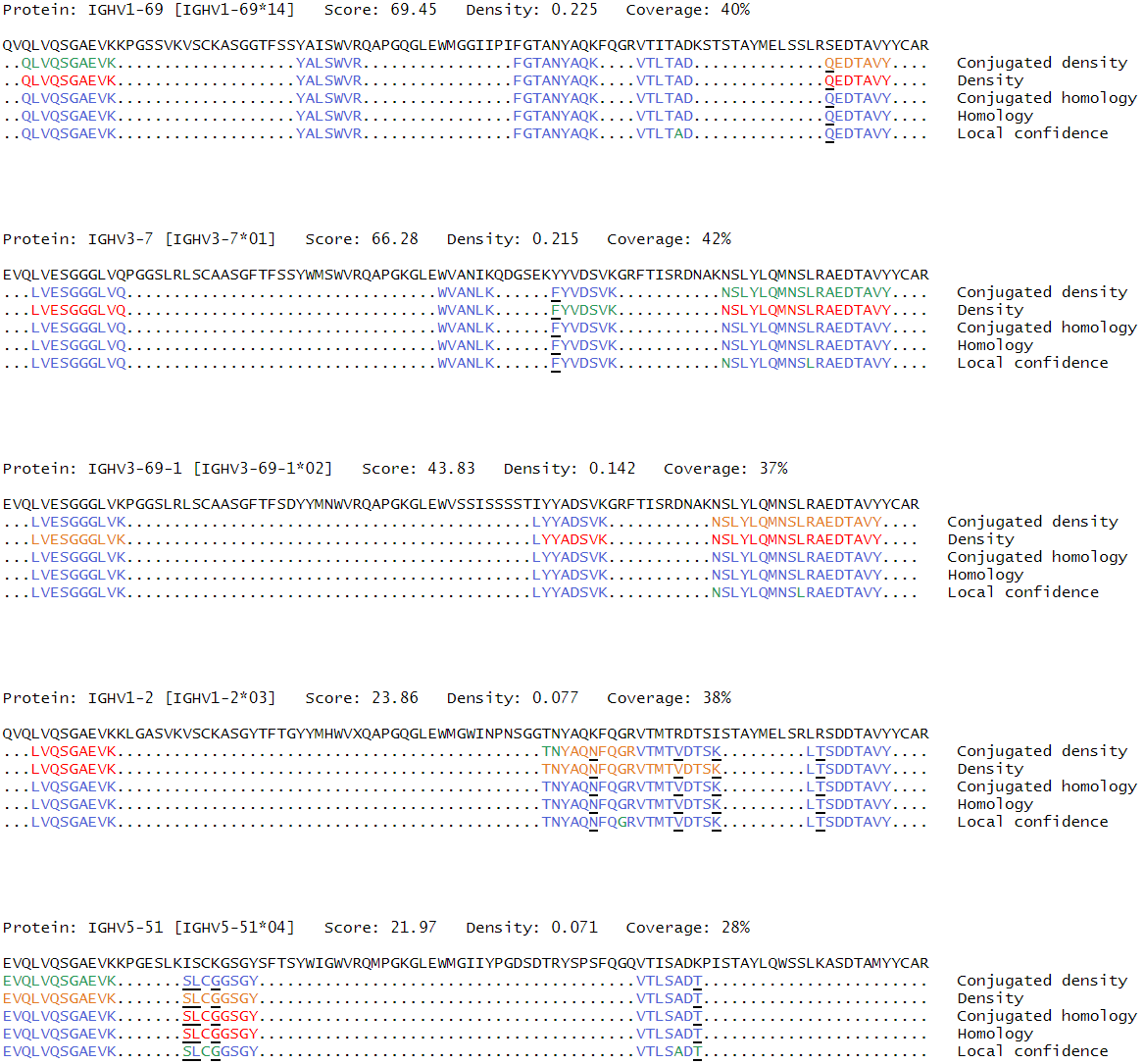


Figure 12 - Gene family summary for file WM16\_B02+B02a\_HAGG\_isolated\_RF\_MJ2.

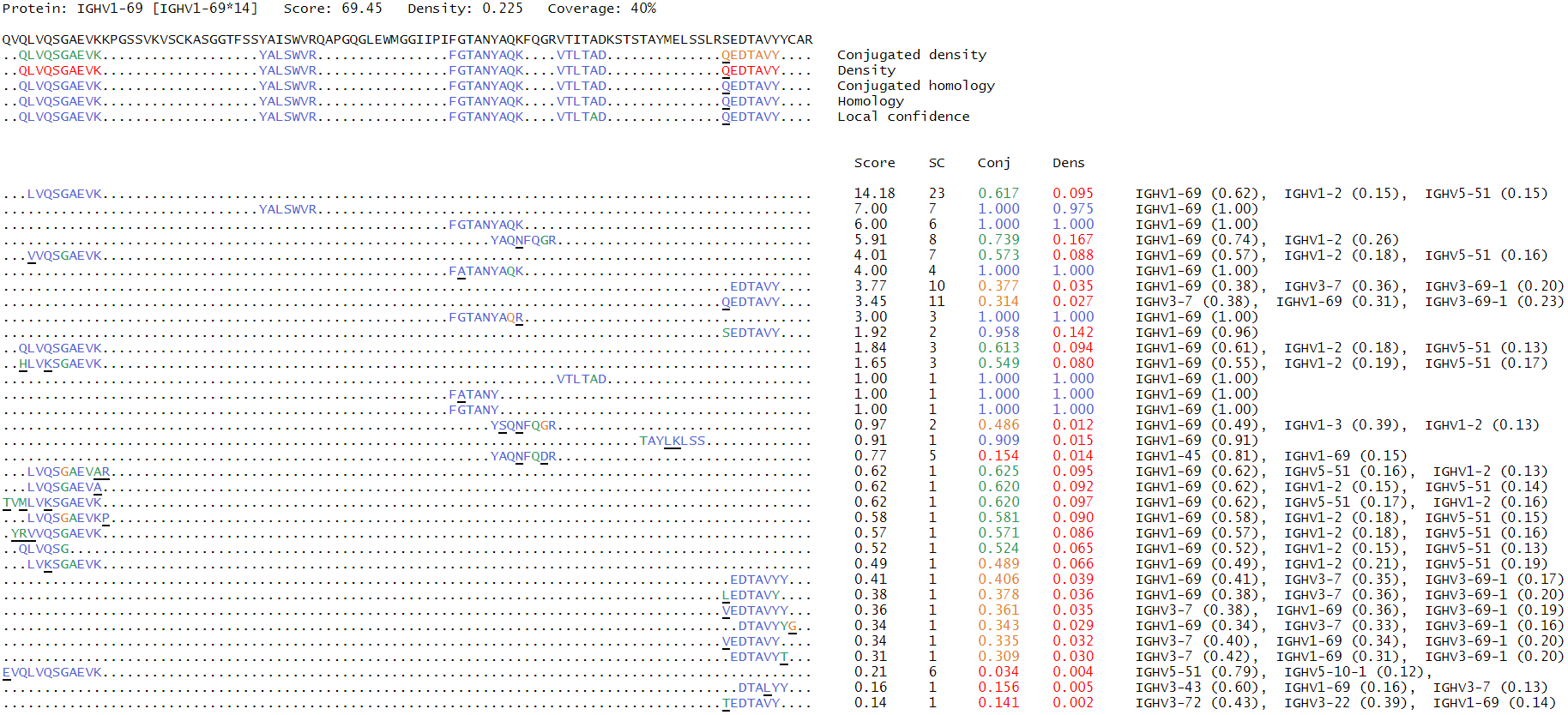


Figure 13 - Gene family #1 from file WM16\_B02+B02a\_HAGG\_isolated\_RF\_MJ2.

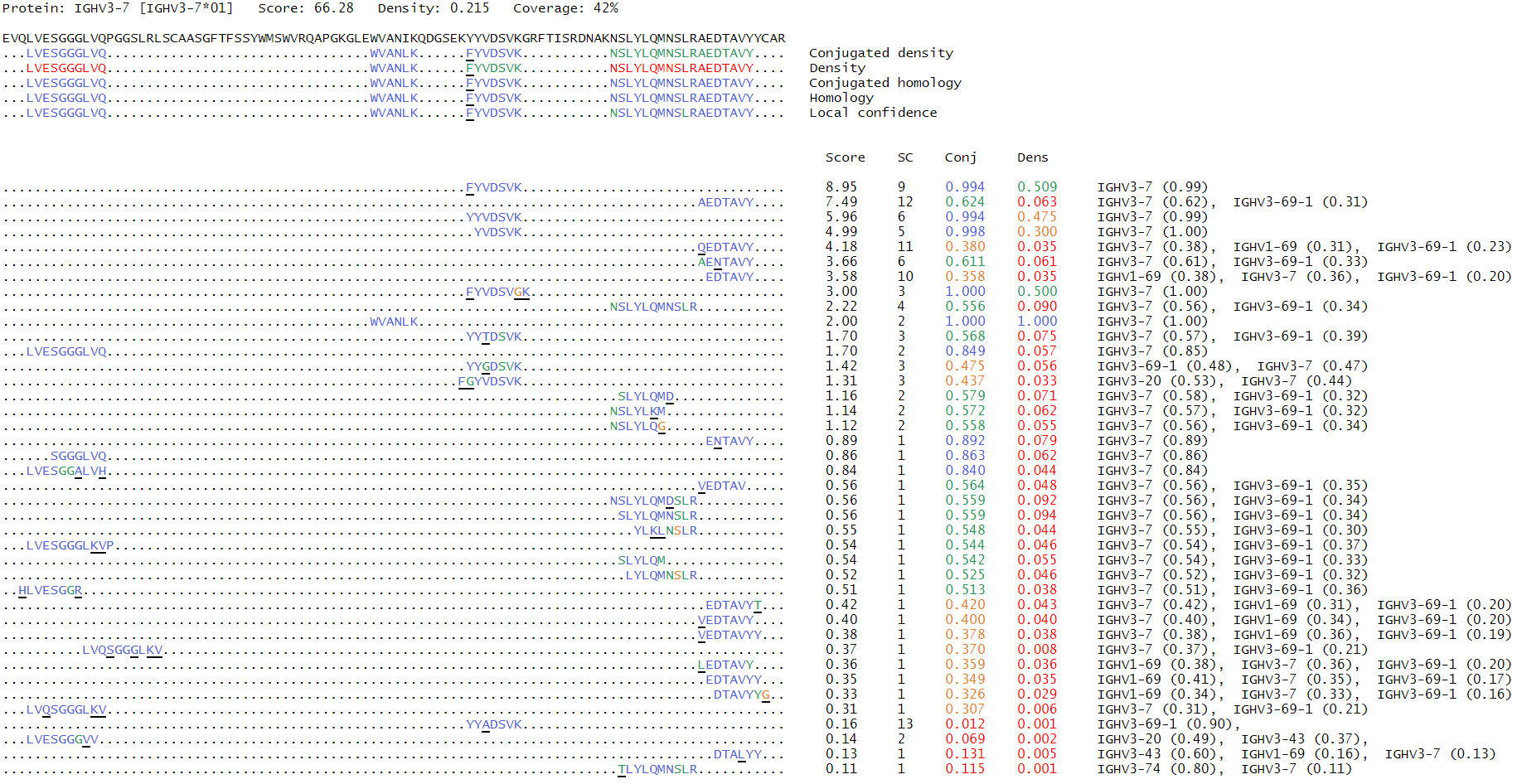


Figure 14 - Gene family #2 from file WM16\_B02+B02a\_HAGG\_isolated\_RF\_MJ2.

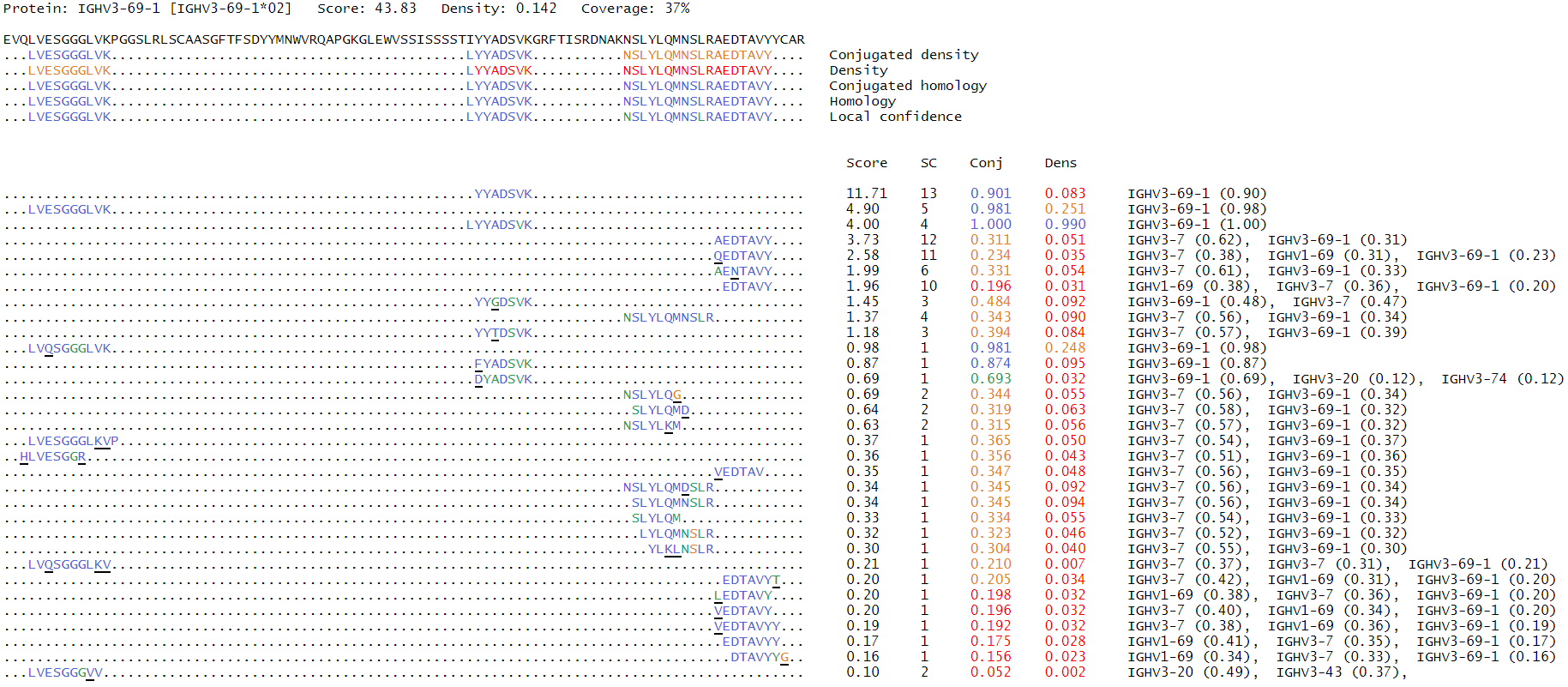


Figure 15 - Gene family #3 from file WM16\_B02+B02a\_HAGG\_isolated\_RF\_MJ2.

## - Where do peptides belong?

The top ranked gene family is IGHV1-69 with a multinomial density of 0.225. An area of interest in the IGHV1-69 is shown in Figure 15. The FGTA motif aligns with the end of the CDR2 with NYAQK following into the FR3. Leading from the CDR2 to the FR3 there is a grouping of peptides that are distinct for IGHV1-69 and support association of shared peptides to this gene family. In particular, the peptide YAQNFQGR was observed with 8 spectra suggesting its sample abundance. Initially shared among IGHV1 and IGHV7 gene families the overall evidence in the sample allowed a stronger association to IGHV1-69 with a increasing from 0.167 to a of 0.739. Such a determination is a combination of robust evidence for IGHV1-69, low evidence for competing assignments, the power of the clustering rate, and the naivity of the model . The importance of considering the weight of evidence and representing this with a conjugation factor is understated. Distinct peptides for IGHV1-69 are centred in a small region and of these YAQNFQGR is the most abundant. It is through conjugation that the role of this peptide in supporting IGHV1-69 is revealed. The peptides YSQNFQGR and YAQNFQDR are also conjugated as a reasonable association to IGHV1-69. However these peptides have both a higher mismatch count of 2 and have a much lower homology density . Although there is still a substantial increase with , the naivity of the model is unable to support association of YSQNFQGR and YAQNFQDR to IGHV1-69 with certainty.

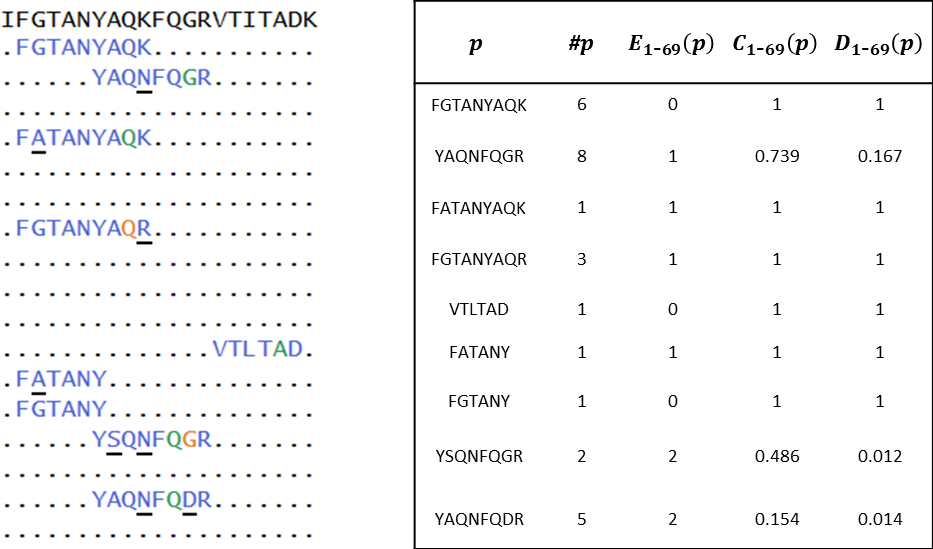


Figure 16 - Peptides from CDR2 into FR3 support gene family IGHV1-69. (a) Alignment example from HTML output. (b) The peptides form a distinct motif for IGHV1-69.

Table 9 - The peptide YAQNFQGR is distributed among IGHV1 and IGHV7 gene families. Evidence in the sample supports an assignment to IGHV1-69 with a increasing from 0.167 to a of 0.739.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| YAQNFQGR | YAQKFQGR | IGHV1-69 | 1 | 0.16742 | 0.73874 |
| YAQNFQGR | YAQKFQGR | IGHV1-2 | 1 | 0.20337 | 0.25911 |
| YAQNFQGR | YAQGFTGR | IGHV7-81 | 2 | 0.00161 | 0.00082 |
| YAQNFQGR | YAQGFTGR | IGHV7-4-1 | 2 | 0.00128 | 0.00059 |
| YAQNFQGR | YAQKFQDR | IGHV1-45 | 2 | 0.00144 | 0.00046 |
| YAQNFQGR | YAEKFQGR | IGHV1-69-2 | 2 | 0.00273 | 0.00016 |
| YAQNFQGR | YSQKFQGR | IGHV1-3 | 2 | 0.00223 | 8.97E-05 |
| YAQNFQGR | YAQKFQGR | IGHV1-24 | 1 | 0.24452 | 2.05E-33 |
| YAQNFQGR | YAQKFQGR | IGHV1-46 | 1 | 0.16742 | 2.31E-43 |
| YAQNFQGR | YAQKFQGR | IGHV1-8 | 1 | 0.20337 | 1.11E-55 |
| YAQNFQGR | YAKKFQGR | IGHV1-68 | 2 | 0.00161 | 8.21E-69 |
| YAQNFQGR | YAQKLQGR | IGHV1-18 | 2 | 0.00180 | 3.31E-149 |
| YAQNFQGR | YAQKFQER | IGHV1-58 | 2 | 0.00113 | 9.51E-187 |

It is noteworthy that following conjugation the peptides associated with a gene family become clearer. Here the grouping of peptides in the CDR2 and the FR3 have almost surely originated from IGHV1-69. As a consequence the variation of these peptides from the germline (which may not necessarily be represented in the database) becomes apparent. An initial observation suggests that FGTANYAQK is a likely progenerator peptide, diverging with FATANYAQK and FGTANYAQR. There is an important consideration in that the spectra that are observed are only those possible by physical limitations of the mass spectrometer. For example, the peptide grouping YAQNFQGR, YSQNFQGR, and YAQNFQDR are probably a result of clonal divergence, however these peptides would only be possible with both a residue change flanking the CDR2 and at the end of the FATANYAQK motif peptides - say, as FGTARYAQK. Any earlier divergences would not be observable: The FGTA peptide produced by a flanking CDR2 substitution would create too small a peptide and neither would the FQGR of the latter group produce spectra. Table 10 proposes a possible clonal drift and states those peptide that would be observable with a trypsin digest.

Table 10 - Clonal divergence of the IGHV1-69 germline could explain the drift of observed peptides. Only the trypsin digested peptide fragments are shown. The proposed divergence is on the basis of the least overall substitutions that could explain the observed peptides.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Peptide | Potential origin | Proposed substitution | Observed? | Spectral count |
| FGTANYAQK | Progenator | - | Y | 6 |
| FQGR | Progenator | - | N | 0 |
| FGTAK | FGTANYAQK | N -> K | N | 0 |
| FGTAR | FGTANYAQK | N -> R | N | 0 |
| YAQK | FGTANYAQK | N -> K / R | N | 0 |
| FGTANY | FGTANYAQK | A -> K / R | Y | 1 |
| FATANY | FGTANY | G -> A | Y | 1 |
| FATANYAQK | FGTANYAQK | G -> A | Y | 1 |
| FGTANYAQR | FGTANYAQK | K -> R | Y | 3 |
| FGTANYAQNFQGR | FGTANYAQK | K -> N | N | 0 |
| YAQNFQGR | FGTANYAQNFQGR | N -> K / R | Y | 8 |
| YSQNFQGR | YAQNFQGR | A -> S | Y | 2 |
| YAQNFQGR | YAQNFQGR | N -> K | Y | 5 |

The gene family with the second highest multinomial density is IGHV3-7 with a of 0.215. There is a region covering the FR2, CDR2, and FR3 that support IGHV3-7. However, unlike the distinct peptides for IGHV1-69, the values of suggests that there is at least one other gene family that could explain the obseved peptides. Analysis through the model values the peptides FYVDSVK, YYVDSVK, and YVDSVK strongly for IGHV3-7 with an increase of from 0.509, 0.475, and 0.300 to a of 0.994, 0.994, and 0.998 respectively. The weight of the evidence of IGHV3-7 also supports the peptides YYTDSVK, YYGDSVK, and FGYVDSVK, although there remains a significant uncertainty of association. The uncertainty is a consequence of a competiting gene family, IGHV3-69-1, that will be described shortly.

Similarly to peptide associations for IGHV1-69 a clonal divergence is apparent. By considering the substitutions necessarily it seems more reasonable that the peptides FYVDSVGK and FGYVDSVK are misassignments of the spectra and contain post-translational modifications rather than amino acids.



Figure 17 - Peptides covering the FR2, CDR2, and FR3 regions support gene family IGHV3-7.

Table 11 - Clonal divergence of the IGHV3-7 germline could explain the drift of observed peptides. Only the trypsin digested peptide fragments are shown. The proposed divergence is on the basis of the least overall substitutions that could explain the observed peptides. The peptides FYVDSVGK and FGYVDSVK are considered to be misassignments of the spectra and more likely to be post-translational modifications than amino acids.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Peptide | Potential origin | Proposed substitution | Observed? | Spectral count |
| YYVDSVK | Progenator | - | Y | 6 |
| FYVDSVK | YYVDSVK | Y -> F | Y | 9 |
| YVDSVK | YYVDSVK | Y -> K / R | Y | 5 |
| YYTDSVK | YYVDSVK | V -> T | Y | 3 |
| YYGDSVK | YYVDSVK | V -> G | Y | 3 |
| YYADSVK | YYVDSVK | V -> A | Y | 13 |

The third ranked gene family is IGHV3-69-1 with a multinomial densityof 0.142. Peptides supporting IGHV3-69-1 cover a narrow region at the end of the CDR2 leading into the FR3. The IGHV3-69-1 gene family has scored well considering the small coverage of significant peptides. This is largely a result of the relatively high spectral count of the distinct peptides YYADSVK and LYYADSVK. The peptide LYYADSVK was initially distinct with a of 0.990 increasing to an almostcertain of1.000. However, the peptide YYADSVK recieves a seemingly disproportionate increase of from 0.083 to a of0.901. In fact, the result of this is that the majority of the score for IGHV3-69-1 is due to the high spectral count of YYADSVK and its *a posteriori* distinctiveness. This seems reasonable, at least if the peptide LYYADSVK could be known to also originate from IGHV3-69-1: all that would be needed to create an observable YYADSVK peptide is a substition of an I residue to a K / R residue. Supposing that the LLYADSVK and YYADSVK peptides did indeed originate from IGHV3-69-1, the proposed clonal divergence in Table 12 is suggested.

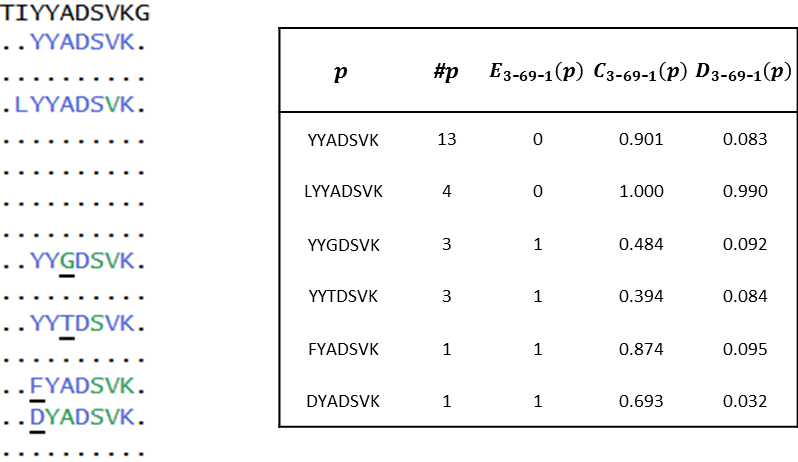


Figure 18- Peptides covering the CDR2 and FR3 support gene family IGHV3-69-1.

Table 12 - Clonal divergence of the IGHV3-69-1 germline could explain the drift of observed peptides. Only the trypsin digested peptide fragments are shown. The proposed divergence is on the basis of the least overall substitutions that could explain the observed peptides.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Peptide | Potential origin | Proposed substitution | Observed? | Spectral count |
| LYYADSVK | Progenator | - | Y | 4 |
| YYADSVK | LYYADSVK | I -> K / R | Y | 13 |
| YYGDSVK | YYADSVK | A -> G | Y | 3 |
| YYTDSVK | YYADSVK | A -> T | Y | 3 |
| FYADSVK | YYADSVK | Y -> F | Y | 1 |
| DYADSVK | YYADSVK | Y ->D | Y | 1 |

## - From proteins to peptides

The clonal divergence of the LYYADSVK peptide form a IGHV3-69-1 germline is suggested to produce a subsequent germline with YYADSVK, which would seem to become the dominant clonal type. There is a difficult problem here, the peptides YYGDSVK and YYTDSVK could also be a result of a IGHV3-7 germline YYVDSVK substitution and this is supported by competiting values of . In contrast the peptides FYADSVK and DYADSVK only require a single substitution from a IGHV3-69-1 germline YYADSVK, but require two substitutions from a IGHV3-7 germline YYVDSVK. Table 13 demonstrates the values of between the IGHV3-7 and IGHV3-69-1 gene families.

Table 13 - Values of for the IGHV3-7 and IGHV3-69-1 gene families. The peptides YYGDSVK and YYTDSVK are not able to be resolved to a distinct origin.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| YYVDSVK | 6 | 0.994 | 0.006 |
| FYVDSVK | 9 | 0.994 | 0.005 |
| YVDSVK | 5 | 0.998 | 0.000 |
| YYTDSVK | 3 | 0.568 | 0.394 |
| YYGDSVK | 3 | 0.475 | 0.484 |
| YYADSVK | 13 | 0.012 | 0.901 |
| LYYADSVK | 4 | 0.000 | 1.000 |
| FYADSVK | 1 | 0.013 | 0.874 |
| DYADSVK | 1 | 0.000 | 0.693 |

The values of suggest that the grouping of peptides is able to be resolved excepting the two peptides YYGDSVK and YYTDSVK. This is not disastarous at least, the majority of peptides are able to be confidently associated to a gene family and the areas of overlap make sense with the substitution rate implicit in the model. However, it could also be argued that such peptides originated entirely from IGHV3-7 or IGHV3-69-1 and that the rate of substitution is much greater. Alternatively, substitutions may be altogether quite unlikely.

## - The problem with parameters

Without a regression model to best fit the parameters, the selection of suitable values is a subjective task, even guided by intuitive results. In general there exists at least one set of parameters that optimally represents that data while also being bound by the constraints of the model. With a full probability model the data could be best fit by regressing towards a performance measure - such as a maximum entropy definition of clustering. However, as the current model has distributed values only for the initial homology scoring, and even still without variance bounding, a maximum entropy model can only be considered through repeat testing. Looking at the effects of adjusting each of the parameters in turn is a simple way of investigating this. Keep in mind that confounding interaction effects are not immediately obvious.

Table 14 - The values

|  |  |  |
| --- | --- | --- |
| Parameter | Definition | Default value |
|  | The parameter value for the weight of the homology score. An increase in this value places more precidence on the relative differences in homology as assigned by BLAST. | 3.5 |
|  | The parameter value for | 0.3 |
|  |  | 0.5 |
|  |  | 1 + (0.005 \* iteration) |
|  |  | 0.005 |

Table 15 - Effect of adjusting parameter .

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| FGTANYAQK | FGTANYAQK | IGHV1-69 | 0 | 0 | 0.99997 | 0.99999 | 0.99999 |
| FGTANYAQK | GNTNYAQK | IGHV1-18 | 3 | 1 | 6.89E-06 | 5.94E-08 | 5.12E-10 |
| FGTANYAQK | GNTNYAQK | IGHV1-45 | 3 | 1 | 6.89E-06 | 5.94E-08 | 5.12E-10 |
| FGTANYAQK | GGTNYAQK | IGHV1-2 | 3 | 1 | 6.33E-06 | 5.27E-08 | 4.39E-10 |
| FGTANYAQK | GNTNYAQK | IGHV1-58 | 3 | 1 | 6.33E-06 | 5.27E-08 | 4.39E-10 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| YALSWVR | YAISWVR | IGHV1-69 | 0 | 0 | 0.91136 | 0.97527 | 0.99297 |
| YALSWVR | YAMSWVR | IGHV3-23 | 1 | 0 | 0.04492 | 0.01442 | 0.00440 |
| YALSWVR | YAMSWVR | IGHV3-47 | 1 | 0 | 0.03316 | 0.00943 | 0.00255 |
| YALSWVR | YAMHWVR | IGHV1-3 | 2 | 0 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMHWVR | IGHV3-9 | 2 | 0 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMHWVR | IGHV3-30-3 | 2 | 0 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMHWVR | IGHV3-43 | 2 | 0 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMSWFR | IGHV3-49 | 2 | 0 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMHWVR | IGHV3-64 | 2 | 0 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMNWVR | IGHV7-4-1 | 2 | 0 | 0.00150 | 0.00012 | 9.77E-06 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| YYVDSVK | YYVDSVK | IGHV3-7 | 0 | 0 | 0.41633 | 0.47473 | 0.49302 |
| YYVDSVK | YYVDSVK | IGHV3-52 | 0 | 0 | 0.41633 | 0.47473 | 0.49302 |
| YYVDSVK | HYVDSVK | IGHV3-16 | 1 | 0 | 0.01763 | 0.00567 | 0.00166 |
| YYVDSVK | YYADSVK | IGHV3-23 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-30 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-30-3 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-30-5 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-33 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-43 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-53 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-66 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-69-1 | 1 | 0 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | IGHV3-64 | 1 | 0 | 0.01378 | 0.00402 | 0.00106 |
| YYVDSVK | HYADSVK | IGHV3-35 | 2 | 0 | 0.00062 | 5.26E-05 | 4.05E-06 |
| YYVDSVK | YADSVK | IGHV3-20 | 2 | 1 | 0.00010 | 4.10E-06 | 1.52E-07 |
| YYVDSVK | YYADSV | IGHV3-47 | 2 | 1 | 9.15E-05 | 3.59E-06 | 1.28E-07 |

Table 16 - Effect of adjusting parameter .

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| FGTANYAQK | FGTANYAQK | IGHV1-69 | 0 | 0 | 0.99999 | 0.99999 | 0.99995 |
| FGTANYAQK | GNTNYAQK | IGHV1-18 | 3 | 1 | 5.81E-13 | 5.94E-08 | 1.27E-05 |
| FGTANYAQK | GNTNYAQK | IGHV1-45 | 3 | 1 | 5.81E-13 | 5.94E-08 | 1.27E-05 |
| FGTANYAQK | GGTNYAQK | IGHV1-2 | 3 | 1 | 5.81E-13 | 5.27E-08 | 1.13E-05 |
| FGTANYAQK | GNTNYAQK | IGHV1-58 | 3 | 1 | 5.16E-13 | 5.27E-08 | 1.13E-05 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| YALSWVR | YAISWVR | IGHV1-69 | 0 | 0 | 0.99948 | 0.97527 | 0.84885 |
| YALSWVR | YAMSWVR | IGHV3-23 | 1 | 0 | 0.00032 | 0.01442 | 0.07503 |
| YALSWVR | YAMSWVR | IGHV3-47 | 1 | 0 | 0.00020 | 0.00943 | 0.04906 |
| YALSWVR | YAMHWVR | IGHV1-3 | 2 | 0 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMHWVR | IGHV3-9 | 2 | 0 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMHWVR | IGHV3-30-3 | 2 | 0 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMHWVR | IGHV3-43 | 2 | 0 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMSWFR | IGHV3-49 | 2 | 0 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMHWVR | IGHV3-64 | 2 | 0 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMNWVR | IGHV7-4-1 | 2 | 0 | 5.83E-08 | 0.00012 | 0.00387 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| YYVDSVK | YYVDSVK | IGHV3-7 | 0 | 0 | 0.49943 | 0.47473 | 0.37879 |
| YYVDSVK | YYVDSVK | IGHV3-52 | 0 | 0 | 0.49943 | 0.47473 | 0.37879 |
| YYVDSVK | HYVDSVK | IGHV3-16 | 1 | 0 | 0.00013 | 0.00567 | 0.02708 |
| YYVDSVK | YYADSVK | IGHV3-23 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-30 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-30-3 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-30-5 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-33 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-43 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-53 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-66 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-69-1 | 1 | 0 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | IGHV3-64 | 1 | 0 | 9.05E-05 | 0.00402 | 0.01919 |
| YYVDSVK | HYADSVK | IGHV3-35 | 2 | 0 | 2.53E-08 | 5.26E-05 | 0.00149 |
| YYVDSVK | YADSVK | IGHV3-20 | 2 | 1 | 1.97E-09 | 4.10E-06 | 0.00012 |
| YYVDSVK | YYADSV | IGHV3-47 | 2 | 1 | 1.73E-09 | 3.59E-06 | 0.00010 |

Table 17 - Effect of adjusting parameter .

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| FGTANYAQK | FGTANYAQK | IGHV1-69 | 0 | 0 | 0.99999 | 0.99999 | 0.99999 |
| FGTANYAQK | GNTNYAQK | IGHV1-18 | 3 | 1 | 9.94E-09 | 5.94E-08 | 2.91E-07 |
| FGTANYAQK | GNTNYAQK | IGHV1-45 | 3 | 1 | 9.94E-09 | 5.94E-08 | 2.91E-07 |
| FGTANYAQK | GGTNYAQK | IGHV1-2 | 3 | 1 | 8.83E-09 | 5.27E-08 | 2.56E-07 |
| FGTANYAQK | GNTNYAQK | IGHV1-58 | 3 | 1 | 8.83E-09 | 5.27E-08 | 2.25E-07 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| YALSWVR | YAISWVR | IGHV1-69 | 0 | 0 | 0.97527 | 0.97527 | 0.97527 |
| YALSWVR | YAMSWVR | IGHV3-23 | 1 | 0 | 0.01442 | 0.01442 | 0.01442 |
| YALSWVR | YAMSWVR | IGHV3-47 | 1 | 0 | 0.00943 | 0.00943 | 0.00943 |
| YALSWVR | YAMHWVR | IGHV1-3 | 2 | 0 | 0.00012 | 0.00012 | 0.00012 |
| YALSWVR | YAMHWVR | IGHV3-9 | 2 | 0 | 0.00012 | 0.00012 | 0.00012 |
| YALSWVR | YAMHWVR | IGHV3-30-3 | 2 | 0 | 0.00012 | 0.00012 | 0.00012 |
| YALSWVR | YAMHWVR | IGHV3-43 | 2 | 0 | 0.00012 | 0.00012 | 0.00012 |
| YALSWVR | YAMSWFR | IGHV3-49 | 2 | 0 | 0.00012 | 0.00012 | 0.00012 |
| YALSWVR | YAMHWVR | IGHV3-64 | 2 | 0 | 0.00012 | 0.00012 | 0.00012 |
| YALSWVR | YAMNWVR | IGHV7-4-1 | 2 | 0 | 0.00012 | 0.00012 | 0.00012 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| YYVDSVK | YYVDSVK | IGHV3-7 | 0 | 0 | 0.47474 | 0.47473 | 0.47473 |
| YYVDSVK | YYVDSVK | IGHV3-52 | 0 | 0 | 0.47474 | 0.47473 | 0.47473 |
| YYVDSVK | HYVDSVK | IGHV3-16 | 1 | 0 | 0.00568 | 0.00567 | 0.00568 |
| YYVDSVK | YYADSVK | IGHV3-23 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-30 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-30-3 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-30-5 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-33 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-43 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-53 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-66 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-69-1 | 1 | 0 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | IGHV3-64 | 1 | 0 | 0.00402 | 0.00402 | 0.00402 |
| YYVDSVK | HYADSVK | IGHV3-35 | 2 | 0 | 5.26E-05 | 5.26E-05 | 5.26E-05 |
| YYVDSVK | YADSVK | IGHV3-20 | 2 | 1 | 6.85E-07 | 4.10E-06 | 1.33E-05 |
| YYVDSVK | YYADSV | IGHV3-47 | 2 | 1 | 6.00E-07 | 3.59E-06 | 1.17E-05 |

It is noteworthy that the more initially distinct a peptide is for a gene family association the less impact there is with a change of parameter. This makes sense: a peptide that is distinctly representative of a protein will be less amenable to variance of the model. In constrast peptides that are heavily shared have in general a smaller proportion of protein association evidence and are more sensitive to the model. It may seem reasonable to focus the inference of the data to those peptides that are most likely to be robust with variance. Indeed, strongly distinct peptides support their associated proteins under a wide distribution of model parameters and are not as likely to succumb to spurious subjectivity. On the other hand, there is good reason to make use of the full body of data, if not only because there is much more of it. Distinct peptides themselves are susceptable to the effect of substitution and the implicit rate of which can result in a loss in what is observed. More importantly, it is critical to determine when a shared peptide acquires a substitution that gives it a distinct façade.

## - A distinct peptide by any other name

Distinct peptides can provide powerful evidence towards a protein of origin. The canonical idea of a distinct peptide often involves determining peptides that have aligned sequence identity with a unique protein counterpart. Such a peptide is often referred to as a *unique peptide*. Although unique peptides have been fundamental in protein mass spectrometry studies, the sequence similarity among the immunoglobulin gene families brings new challenges.

Consider the peptide FGTANYAQK shown in Table 18. This is almost surely a unique peptide for IGHV1-69.

Table 18 - The IGHV1-69 unique peptide FGTANYAQK.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| FGTANYAQK | FGTANYAQK | IGHV1-69 | 0 | 0 | 0.99999 |
| FGTANYAQK | GNTNYAQK | IGHV1-18 | 3 | 1 | 5.94E-08 |
| FGTANYAQK | GNTNYAQK | IGHV1-45 | 3 | 1 | 5.94E-08 |
| FGTANYAQK | GGTNYAQK | IGHV1-2 | 3 | 1 | 5.27E-08 |
| FGTANYAQK | GNTNYAQK | IGHV1-58 | 3 | 1 | 5.27E-08 |

## - Distance is relative

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| query | subject | subject\_accession | mismatch\_count | homology\_density |
| YYVDSVK | YYVDSVK | IGHV3-7 | 0 | 0.472811 |
| YYVDSVK | YYVDSVK | IGHV3-52 | 0 | 0.472811 |
| YYVDSVK | HYVDSVK | IGHV3-16 | 1 | 0.00565536 |
| YYVDSVK | YYADSVK | IGHV3-53 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-66 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-23 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-43 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-30-5 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-30 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-30-3 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-69-1 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-33 | 1 | 0.00451191 |
| YYVDSVK | YYADSVK | IGHV3-64 | 1 | 0.00400709 |
| YYVDSVK | YYADSVK | IGHV3-48 | 1 | 0.00400709 |
| YYVDSVK | HYADSVK | IGHV3-35 | 2 | 5.24E-05 |
| YYVDSVK | YADSVK | IGHV3-20 | 2 | 2.61E-05 |
| YYVDSVK | YYADSV | IGHV3-47 | 2 | 2.29E-05 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| query | subject | subject\_accession | mismatch\_count | homology\_density |
| YYADSVK | YYADSVK | IGHV3-23 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-53 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-64 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-48 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-66 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-21 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-43 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-69-1 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-30-5 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-30 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-30-3 | 0 | 0.0829693 |
| YYADSVK | YYADSVK | IGHV3-33 | 0 | 0.0829693 |
| YYADSVK | HYADSVK | IGHV3-35 | 1 | 0.000888038 |
| YYADSVK | YYVDSVK | IGHV3-52 | 1 | 0.000791754 |
| YYADSVK | YADSVK | IGHV3-20 | 1 | 0.000448287 |
| YYADSVK | YYVDSVK | IGHV3-7 | 1 | 0.000703167 |
| YYADSVK | YYADSV | IGHV3-47 | 1 | 0.00039813 |
| YYADSVK | YADSVK | IGHV3-74 | 1 | 0.00039813 |
| YYADSVK | YADSVK | IGHV3-9 | 1 | 0.000352115 |
| YYADSVK | YADSVK | IGHV3-11 | 1 | 0.000352115 |
| YYADSVK | HHADSVK | IGHV3-32 | 2 | 8.10E-06 |
| YYADSVK | HYVDSVK | IGHV3-16 | 2 | 7.10E-06 |
| YYADSVK | HHADSVK | IGHV3-29 | 2 | 7.10E-06 |
| YYADSVK | HHADSVK | IGHV3-30-42 | 2 | 7.10E-06 |
| YYADSVK | HHADSVK | IGHV3-30-22 | 2 | 7.10E-06 |

# Future

## - Analysis of data generation and reproduction

It is observed that spectral reproduction from sequential samples are reasonably consistent. Assigned peptides from an earlier sample are often seen in a later sample with predictable levels of corresponding spectra. However, the reproduction is not identical, and variance is seen from run to run. In particular the reproduction of spectra over extended periods of time has not been well studied. Claims on the basis of spectral assignment are reliant on the consistency of independent data generation. It could be valuable to study the production of spectra over subsequent and seperated analyses. Peptide assignments could be compared from associated runs, or the generation of the spectra themselves using fragment ion matching. The production of spectra may differ depending on the spatial region of the originating peptide.

## - Analysis of germline divergence

The clonal divergence of germline immunoglobulins is readily seen in the collection of assigned peptides.

## - Blastp internal integration

There are notable limitations with using blastp through an external command line. Although parameters are able to be directed to blastp during IgFamily runtime, the modifiable parameters, while thorough, are restricted. In particular blastp sets a hard-coded threshold for score output. This would not be an oversight for simple protein identification, but it would be appropriate to build a complete model that assigns some probability to peptide association for each protein. The effect of this may not be significant except through certain cases of conjugation but would also serve to satisfy the formal definition of the model.

## - Blastp custom substitution matrix

The BLOSUM62 substitution matrix described in the Story is most often used for defining the rate of amino acid substitution. Specifically it describes the relative likelihood of any amino acid change occurring against an expected query amino acid. Although the BLOSUM62 is routinely used for determining sequence similarity, it is may not represent the rate of residue change in immunoglobulins. By modelling the certainty of peptide assignment, such as with the method proposed in this reading, the divergence from the germline can be used to construct a residue substitution frequency matrix. This could better represent observed clonal divergence when used for analysis. Note that the frequency matrix would necessarily be created by the frequency of observed peptides, and the observation of these peptides is biased by the peptides produced by a particular digest. This would need to considered into the frequency model for greater specificity.

## - Blastp custom conservation weighting

The BLOSUM62 substitution matrix described in the Story above assumes that the rate of substitution is spatially uniform - That is, any amino acid is equally likely to be substituted than any other through the substition matrix. It is known that this is not the case for immunoglobulins, where a relatively greater likelihood of mutation occurs in the hypervariable regions.

## - Advanced statistical modellng

## - User interface and interaction

The *IgFamily* program is convenient as a command line application. The algorithmic process is largely linear in process and user intervention is often not required.

## - Data filesystem

## - Further automation

## - De novo and database process comparison

The algorithmic differences between de novo and database peptide assignment are suspected to confer some level of error to spectral assignment. It is not well understood what factors are responsible for incorrect assignment in either case. The IgFamily stores information relating scan number to peptide assignment. With this the de novo and database assignments are able to be compared in each method. It would be worthwhile to determine which conditions are likely to produce an incorrect assignment.

## - De novo and database process combination

There are instances of de novo assignment when a peptide generates a low confidence assignment but is also supported by the same assignment through the database method. It is possible to compare instances when this occurs and support low confidence de novo assignments. This can be extended by also considering commonly observed misassignments. For example, often the termal amino acid assignments are incorrect by having the residue locations reversed

## - Contaminants report and exclusion list generation

The IgFamily program is able to recognise peptides that originate from contaminant proteins. From these a report is able to generated with the peptide, mass-to-charge ratio, retention time, charge, theoretical mass, and the associated contaminant protein. In addition exclusion list is produced in a format suitable for the AB Sciex mass spectrometry software. Reports and exclusion lists are able to be generated from multiple samples by combining the results or by selecting only those peptides that are observed in a proportion of the samples.

# Production

Version: v0.12.3

Release: 2016-09-17

Language: C++14

Codebase: 4,857 source lines of code

Dependencies: 1 .cpp file, 16 .h files, msconvert.exe, NOVOR.bat, makeblastdb.exe, blastp.exe

Development environment: Microsoft Visual Studio Community 2015

Version control: Git

Version history: 139 commits

Codebase additions: 21,040 source lines of code

Codebase deletions: 16,083 source lines of code

# References

Frost, S. et al., 2015. Assigning and visualizing germline genes in antibody repertoires. *Philosophical Transcactions of the Royal Society B.*

Karlin, S. & Altschul, S., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences,* Volume 87, pp. 2264-2268.

# Workshop

“*Prior, prior, pants on fire.* Historically, some opponents of Bayensian inference objected to the arbitrariness of prior distributions. It’s true that priors are very flexible, being able to encode many different states of information. If the prior can be anything, isn’t it possible to get any answer you want? Indeed it is. Regardless, after a couple hundred years of Bayesian calculation, it hasn’t turned out that people use priors to lie. If your goal is to lie with statistics, you’d be a fool to do it with priors, because such a lie would easily be uncovered. Better to use the more opaque macinery of the likelihood. Or better yet - don’t actually take this advice! - massage the data, drop some “outliers”, and otherwise engage in motivated data transformation.

It is true enough that choice of likelihood is much more conventionalised than choice of prior. But conventional choices are often poor ones, smuggling in influences that can be hard to discover. In this regard, both Bayesian and non-Bayesian models are equally harried, because both traditions depend heavily upon likelihood functions and conventionalised model forms. And the fact that the non-Bayesian procedure doesn’t have to make an assumption about the prior is of little comfort. This is because non-Bayesian procedures need to make choices that Bayesian ones do not, such as choice of estimator or likelihood penalty. Often, such choices can be shown to be equivalent to some Bayesian choice of prior.”

“A basic issue with interpreting model-based estimates is in knowing the meaning of parameters. There is no consensus about what a parameter means, however, because different people take different philosophical stances towards models. The perspective in this book is a common Bayesian perspective: Posterior probabilities of parameter values describe the relative compatibility of different states of the world with the data, according to the model. These are small world numbers. Reasonable people may disagree with the large world meaning, and the details of those disagreements depend strongly upon context. Such disagreements are productive, because they lead to model criticism and revision.”