IgFamily v0.12.2e

Technical manual v0.3.8

Preface

Mass spectrometry immunoproteomics has the potential to investigate serum repetoires as an accurate and high-throughput method. Particularly applicable to patients with immunology maladies, immunoglobulins are able to be selective purified using an antigen or affinity-specific molecule. A sample of immunoglobulins 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. 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 spectra quality, spectral count, supporting peptides, and overall protein coverage.

Although this routine method has proven successful in general studies, there are constraining complications implicit for immunoproteomics. In particular, the comparative database of immunoglobulins is derived from an ancestral repetition of successful germline alleles, referred to as immunoglobulin gene families. There are spatially 80 variable region gene family alleles, each with as many as 20 polymorphic allele variants, with potentially many more not yet described. The variable region has regions of relatively strong conservation and hypervarible regions with a somatically increased nucleotide mutation rate. The generally short-lengthed peptides generated through a mass spectrometry method often results in many peptides corresponding to regions of the database protein that are on one case strongly conserved but on the other prone to excessive mutagenic variation. These confounding factors raise concerns of the validity of current mass spectrometry derived immunoproteomic approaches.

An automated method of inferring from the large bodies of generated data was sought. Initially, the *IgFamily* program was designed to emulate the conventional approach of focusing the inferential effort on representative peptides. As the program was developed interest grew in the role of supporting peptides...

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

Story

**- *Data: what is it?***

Mass spectrometry immunoproteomics 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 for 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 nothing 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. *IgFamily* is able to analyse PEAKS de novo assigned peptides, PEAKS database assigned peptides, and NOVOR de novo assigned peptides.

***- Data: where does it come from?***

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

There exists a distinct conditional probability for each of all peptides to originate from each of all proteins .

The determination of the conditional probability is in general 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 *IgFamily* program are immunoglobulins, and the nature of immunoglobulin diversification creates an almost continuum of possible assignments. 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 particular, 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.



***- 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.

Table - 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, TGLVTVSSASTK, and ASTVLVSSASTK 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. 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*. 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.



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

Although useful as a standalone tool, the BLAST algorithm is designed as an aid for peptide or protein identification. 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.

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 - 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 **P** is sampled, without knowing the peptide sequence, what is the likelihood that it belongs to some protein **G**?

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

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

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

Figure - 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 probability 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 i-th peptide for the k-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 i-th peptide for the k-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 i-th peptide for the k-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 k-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 i-th peptide for the k-th protein and repeat for all peptides and proteins . The transformed homology density is the value of the transformed homology of the i-th peptide for the k-th protein in comparison to the sum of all transformed homologies for the i-th peptide for all proteins -

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

Determine the peptide score of the i-th peptide for the k-th protein and repeat for all peptides and proteins . The peptide score is a product of the conjugated transformed 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 transformed 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 -

Finally, determined the multinomial value for the k-th protein and repeat for all proteins . The multinomial value is the summation of all the peptide scores for all peptides . The multinomial value represents the evidence for the protein as inferred by the model.

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.

Multinomial

Dirichlet

Poisson

Exponential

Figure 3 – 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. The following except from Richard McElreath’s *Statistical Rethinking* exemplifies this:

“*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.”

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".

Examples

To clarify the use of the IgFamily program examples are always necessary. The examples here are generously provided by collaboration -

K01 - A patient with RNP70-positive mixed connective tissue disease. The data is a duplicate sample of serum IgMs that have been purified with the HAGG pulldown method.

K02 - A patient with RNP70-positive mixed connective tissue disease. The data is a duplicate sample of serum IgMs that have been purified with the HAGG pulldown method.

B02 - A patient with primary Sjogrens Syndrome . Here the data is a duplicate sample of serum IgMs that have been purified with the HAGG pulldown method.

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 their density in the mulitnomial distribution while the peptides are ranked by the value . 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 and important information is conveyed. Note that all proteins in the database are considered while only the gene family proteins are shown.

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

|  |  |
| --- | --- |
| Header | Definition |
| Conjugated density | The value of |
| Density | The value of |
| Conjugated homology | The value of |
| Homology | The value of |
|  |  |

Function

*For further information on features, refer to associated entries -*

**- FASTA utility tools:**

Read FASTA files into runtime. Custom FASTA creation utility provides functionality for tailored output files. This utlility can be used to create required FASTA format –

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

Accession field is a housekeeping field and is not required for runtime. Name, type, and species fields are used in data structure creation, association, and function.

|  |  |
| --- | --- |
| FASTA parse function | pp. |
| FASTA file utilities | pp. |

**- Peptide file compatibility:**

Read peptide files into runtime. Currently supported are *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.

*PEAKS v8.0 DE NOVO* de novo peptides peptide file assigns scan number, peptide with modification, and amino acid local confidence score. Note that here *PEAKS* *v8.0 DE NOVO* assigns individual export accessions to replicate peptide assignments.

*PEAKS v8.0 SPIDER* protein peptide file assigns scan number, peptide with modification, spectral count, and -10IgP certainty score.

*NOVOR v1.1* de novo peptides file assigns scan number, peptide with modification, and amino acid local confidence score. Note that here *NOVOR v1.1* assigns individual export accessions to replicate peptide assignments.

|  |  |
| --- | --- |
| Peptide file parse functions | pp. |

**- *msconvert* external integration**

*msconvert* is able to be called through a user defined interface option to convert .wiff and .wiff.scan files 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.

|  |  |
| --- | --- |
| *msconvert* convert command line functions | pp. |
|  |  |

**- *NOVOR v1.1* external integration**

*NOVOR v1.1* is able to be called through a user defined interface option to generate *NOVOR v1.1* de novo peptide 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.

|  |  |
| --- | --- |
| *NOVOR v1.1* convert command line functions | pp. |

**- Local directory runtime 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 *IgFamily* root directory and executes the program. The input file and output files are moved to a folder created in the root directory with the name of the input file sample - The input file contains the sample name and is supported in the three input data types.

|  |  |
| --- | --- |
| Local directory runtimefunctions | pp. |

**- Data filesystem**

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. The user is required to accession a file initially, however there is proposed functionality for dynamic file management with the *ISO/IEC TS 18822:2015* filesystem library.

|  |  |
| --- | --- |
| Data filesystem functions | pp. |

**- Runtime user interface**

On execution of the *IgFamily* program, the user is greeted with an interactive menu. The user is able to access FASTA file utilities, *msconvert* command 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.

|  |  |
| --- | --- |
| User interface functions | pp. |

**- FASTA data structuring**

Following user selection of runtime parameters, the *IgFamily* program will parse the FASTA file(s). The data are stored on contiguous RAM buffers at runtime. 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.

|  |  |
| --- | --- |
| FASTA data structuring functions | pp. |

**- Peptide data structuring**

Following user selection of runtime parameters, the *IgFamily* program will parse the data .csv file(s). The data are stored on contiguous RAM buffers at runtime. 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 of a peptide cleaved at a midpoint such that two peptides of length >= 5 are produced, both are assigned to unique data structures.

|  |  |
| --- | --- |
| Peptide data structuring functions | pp. |

**- External blastp integration**

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 tested against the blast reference database and a results file generated as an output file. The blastp output is parsed into runtime and assigned to a homology data structure. Data assigned are the blastp query, the blastp query alignment, the blastp subject, the blastp subject database accession, the blastp query alignment index, the blastp subject alignment index, and the blastp sequence alignment expectation value. Blastp is programmed to allow up to 200 matches for each query, and a generous threshold for alignment acceptance.

|  |  |
| --- | --- |
| Peptide data structuring functions | pp. |

**- Homology data structuring**

Generated and parsed homology data is runtime associated with respective FASTA and peptide data. Additional data types are created -

- Transformed expectation value: The e-value transformed as a metric of relative likelihood.

- Conjugated expectation value: The e-value transformed to consider overall family evidence.

- Homology parameter density: The density of the e-value (0<=par<=1) compared to all

|  |  |
| --- | --- |
| Peptide data structuring functions | pp. |

**IgFamily v0.12.1h**

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 designed data structures, create blastp input file and blastp database:

7a. Perform create\_blastp\_input().

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().

Table - 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 | 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 | 659.79 | 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 - 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 - 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 | 0.474733 | 0.992769 | 26.2091 |
| 3400 | 2.29E+12 | 8.66E-67 | 0.474733 | 1.27E-78 | 3.36E-77 |
| 3200 | 2.74E+10 | 7.79E-130 | 0.005678 | 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 | 0.000565 | 0.013552 |
| 3000 | 2.19E+10 | 1.09E+07 | 0.00453 | 1.60E-05 | 0.000383 |
| 3000 | 2.19E+10 | 1.09E+07 | 0.00453 | 1.60E-05 | 0.000383 |
| 3000 | 2.19E+10 | 165496 | 0.00453 | 2.43E-07 | 5.83E-06 |

Table - 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 |

Considerations

**- *A unique peptide not by any other name***

Unique peptides can provide strong support for a gene family and have been used in our lab and among the literature for a range of studies. However, there is a difficulty with unique peptides and immunoglobulins, such that the *uniqueness* or *distinctiveness* of a peptide is not necessarily determined by complete peptide sequence identity.

Consider the IGHV1-69 unique peptide shown in Table X from a truncated homology\_summary output.

Table X - The IGHV1-69 unique peptide FGTANYAQK.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| query | subject | subject\_accession | mismatch\_count | homology | homology\_density\_conjugated |
| FGTANYAQK | FGTANYAQK | IGHV1-69 | 0 | 47 | 22256.4 |
| FGTANYAQK | GNTNYAQK | IGHV1-45 | 3 | 19 | 2112.32 |
| FGTANYAQK | GNTNYAQK | IGHV1-18 | 3 | 19 | 2112.32 |
| FGTANYAQK | GNTNYAQK | IGHV1-58 | 3 | 18 | 1835.31 |
| FGTANYAQK | GGTNYAQK | IGHV1-2 | 3 | 18 | 1835.31 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

**- 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 - Selected parameters.

|  |  |
| --- | --- |
| Parameter | Value |
| =PARAMETER\_HOMOLOGY\_WEIGHT | 3.5 |
| =PARAMETER\_HOMOLOGY\_MISMATCH\_WEIGHT | 0.3 |
| =PARAMETER\_HOMOLOGY\_DELTA\_ALIGNMENT\_WEIGHT | 0.5 |
| =PARAMETER\_LOGISTIC\_CONJUGATION\_FACTOR | 1 + (0.005 \* iteration) |
| =PARAMETER\_PRIOR\_DISTRIBUTION\_WEIGHT | 0.005 |

Table - Effect of adjusting parameter .

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| FGTANYAQK | FGTANYAQK | 1-69 | 4700 | 0 | 0 | 1.51E+09 | 7.12E+12 | 3.35E+16 | 0.99997 | 0.99999 | 0.99999 |
| FGTANYAQK | GNTNYAQK | 1-18 | 3000 | 3 | 1 | 10438.5 | 422759 | 1.71E+07 | 6.89E-06 | 5.94E-08 | 5.12E-10 |
| FGTANYAQK | GNTNYAQK | 1-45 | 3000 | 3 | 1 | 10438.5 | 422759 | 1.71E+07 | 6.89E-06 | 5.94E-08 | 5.12E-10 |
| FGTANYAQK | GGTNYAQK | 1-2 | 2900 | 3 | 1 | 9590.24 | 375458 | 1.47E+07 | 6.33E-06 | 5.27E-08 | 4.39E-10 |
| FGTANYAQK | GNTNYAQK | 1-58 | 2900 | 3 | 1 | 9590.24 | 375458 | 1.47E+07 | 6.33E-06 | 5.27E-08 | 4.39E-10 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| YALSWVR | YAISWVR | 1-69 | 3500 | 0 | 0 | 7.25E+08 | 2.54E+12 | 8.88E+15 | 0.91136 | 0.97527 | 0.99297 |
| YALSWVR | YAMSWVR | 3-23 | 3500 | 1 | 0 | 3.57E+07 | 3.75E+10 | 3.94E+13 | 0.04492 | 0.01442 | 0.00440 |
| YALSWVR | YAMSWVR | 3-47 | 3100 | 1 | 0 | 2.64E+07 | 2.45E+10 | 2.28E+13 | 0.03316 | 0.00943 | 0.00255 |
| YALSWVR | YAMHWVR | 1-3 | 3000 | 2 | 0 | 1.20E+06 | 3.23E+08 | 8.73E+10 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMHWVR | 3-9 | 3000 | 2 | 0 | 1.20E+06 | 3.23E+08 | 8.73E+10 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMHWVR | 3-30-3 | 3000 | 2 | 0 | 1.20E+06 | 3.23E+08 | 8.73E+10 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMHWVR | 3-43 | 3000 | 2 | 0 | 1.20E+06 | 3.23E+08 | 8.73E+10 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMSWFR | 3-49 | 3000 | 2 | 0 | 1.20E+06 | 3.23E+08 | 8.73E+10 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMHWVR | 3-64 | 3000 | 2 | 0 | 1.20E+06 | 3.23E+08 | 8.73E+10 | 0.00150 | 0.00012 | 9.77E-06 |
| YALSWVR | YAMNWVR | 7-4-1 | 3000 | 2 | 0 | 1.20E+06 | 3.23E+08 | 8.73E+10 | 0.00150 | 0.00012 | 9.77E-06 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| YYVDSVK | YYVDSVK | 3-7 | 3400 | 0 | 0 | 6.74E+08 | 2.29E+12 | 7.79E+15 | 0.41633 | 0.47473 | 0.49302 |
| YYVDSVK | YYVDSVK | 3-52 | 3400 | 0 | 0 | 6.74E+08 | 2.29E+12 | 7.79E+15 | 0.41633 | 0.47473 | 0.49302 |
| YYVDSVK | HYVDSVK | 3-16 | 3200 | 1 | 0 | 2.86E+07 | 2.74E+10 | 2.63E+13 | 0.01763 | 0.00567 | 0.00166 |
| YYVDSVK | YYADSVK | 3-23 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-30 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-30-3 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-30-5 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-33 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-43 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-53 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-66 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-69-1 | 3000 | 1 | 0 | 2.43E+07 | 2.19E+10 | 1.97E+13 | 0.01500 | 0.00453 | 0.00124 |
| YYVDSVK | YYADSVK | 3-64 | 2900 | 1 | 0 | 2.23E+07 | 1.94E+10 | 1.69E+13 | 0.01378 | 0.00402 | 0.00106 |
| YYVDSVK | HYADSVK | 3-35 | 2800 | 2 | 0 | 1.01E+06 | 2.54E+08 | 6.40E+10 | 0.00062 | 5.26E-05 | 4.05E-06 |
| YYVDSVK | YADSVK | 3-20 | 2700 | 2 | 1 | 162720 | 1.98E+07 | 2.40E+09 | 0.00010 | 4.10E-06 | 1.52E-07 |
| YYVDSVK | YYADSV | 3-47 | 2600 | 2 | 1 | 148069 | 1.73E+07 | 2.03E+09 | 9.15E-05 | 3.59E-06 | 1.28E-07 |

Table - Effect of adjusting parameter .

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| FGTANYAQK | FGTANYAQK | 1-69 | 4700 | 0 | 0 | 1.51E+09 | 7.12E+12 | 7.12E+12 | 0.99999 | 0.99999 | 0.99995 |
| FGTANYAQK | GNTNYAQK | 1-18 | 3000 | 3 | 1 | 10438.5 | 422759 | 9.03E+07 | 0.99999 | 5.94E-08 | 1.27E-05 |
| FGTANYAQK | GNTNYAQK | 1-45 | 3000 | 3 | 1 | 10438.5 | 422759 | 9.03E+07 | 5.81E-13 | 5.94E-08 | 1.27E-05 |
| FGTANYAQK | GGTNYAQK | 1-2 | 2900 | 3 | 1 | 9590.24 | 375458 | 8.02E+07 | 5.81E-13 | 5.27E-08 | 1.13E-05 |
| FGTANYAQK | GNTNYAQK | 1-58 | 2900 | 3 | 1 | 9590.24 | 375458 | 8.02E+07 | 5.16E-13 | 5.27E-08 | 1.13E-05 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| YALSWVR | YAISWVR | 1-69 | 3500 | 0 | 0 | 2.54E+12 | 2.54E+12 | 2.54E+12 | 0.99948 | 0.97527 | 0.84885 |
| YALSWVR | YAMSWVR | 3-23 | 3500 | 1 | 0 | 8.02E+08 | 3.75E+10 | 2.24E+11 | 0.00032 | 0.01442 | 0.07503 |
| YALSWVR | YAMSWVR | 3-47 | 3100 | 1 | 0 | 5.25E+08 | 2.45E+10 | 1.47E+11 | 0.00020 | 0.00943 | 0.04906 |
| YALSWVR | YAMHWVR | 1-3 | 3000 | 2 | 0 | 147885 | 3.23E+08 | 1.16E+10 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMHWVR | 3-9 | 3000 | 2 | 0 | 147885 | 3.23E+08 | 1.16E+10 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMHWVR | 3-30-3 | 3000 | 2 | 0 | 147885 | 3.23E+08 | 1.16E+10 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMHWVR | 3-43 | 3000 | 2 | 0 | 147885 | 3.23E+08 | 1.16E+10 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMSWFR | 3-49 | 3000 | 2 | 0 | 147885 | 3.23E+08 | 1.16E+10 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMHWVR | 3-64 | 3000 | 2 | 0 | 147885 | 3.23E+08 | 1.16E+10 | 5.83E-08 | 0.00012 | 0.00387 |
| YALSWVR | YAMNWVR | 7-4-1 | 3000 | 2 | 0 | 147885 | 3.23E+08 | 1.16E+10 | 5.83E-08 | 0.00012 | 0.00387 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| YYVDSVK | YYVDSVK | 3-7 | 3400 | 0 | 0 | 2.29E+12 | 2.29E+12 | 2.29E+12 | 0.49943 | 0.47473 | 0.37879 |
| YYVDSVK | YYVDSVK | 3-52 | 3400 | 0 | 0 | 2.29E+12 | 2.29E+12 | 2.29E+12 | 0.49943 | 0.47473 | 0.37879 |
| YYVDSVK | HYVDSVK | 3-16 | 3200 | 1 | 0 | 5.86E+08 | 2.74E+10 | 1.64E+11 | 0.00013 | 0.00567 | 0.02708 |
| YYVDSVK | YYADSVK | 3-23 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-30 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-30-3 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-30-5 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-33 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-43 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-53 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-66 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-69-1 | 3000 | 1 | 0 | 4.68E+08 | 2.19E+10 | 1.31E+11 | 0.00010 | 0.00453 | 0.02160 |
| YYVDSVK | YYADSVK | 3-64 | 2900 | 1 | 0 | 2.23E+07 | 1.94E+10 | 1.16E+11 | 9.05E-05 | 0.00402 | 0.01919 |
| YYVDSVK | HYADSVK | 3-35 | 2800 | 2 | 0 | 1.01E+06 | 2.54E+08 | 9.07E+09 | 2.53E-08 | 5.26E-05 | 0.00149 |
| YYVDSVK | YADSVK | 3-20 | 2700 | 2 | 1 | 162720 | 1.98E+07 | 7.06E+08 | 1.97E-09 | 4.10E-06 | 0.00012 |
| YYVDSVK | YYADSV | 3-47 | 2600 | 2 | 1 | 148069 | 1.73E+07 | 6.19E+08 | 1.73E-09 | 3.59E-06 | 0.00010 |

Table - Effect of adjusting parameter .

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| FGTANYAQK | FGTANYAQK | 1-69 | 4700 | 0 | 0 | 7.12E+12 | 7.12E+12 | 5.48E+10 | 0.99999 | 0.99999 | 0.99999 |
| FGTANYAQK | GNTNYAQK | 1-18 | 3000 | 3 | 1 | 70733 | 422759 | 15943.9 | 9.94E-09 | 5.94E-08 | 2.91E-07 |
| FGTANYAQK | GNTNYAQK | 1-45 | 3000 | 3 | 1 | 70733 | 422759 | 15943.9 | 9.94E-09 | 5.94E-08 | 2.91E-07 |
| FGTANYAQK | GGTNYAQK | 1-2 | 2900 | 3 | 1 | 62818.9 | 375458 | 14038.3 | 8.83E-09 | 5.27E-08 | 2.56E-07 |
| FGTANYAQK | GNTNYAQK | 1-58 | 2900 | 3 | 1 | 62818.9 | 375458 | 12301.2 | 8.83E-09 | 5.27E-08 | 2.25E-07 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| YALSWVR | YAISWVR | 1-69 | 3500 | 0 | 0 | 2.54E+12 | 2.54E+12 | 2.54E+12 | 0.99948 | 0.97527 | 0.97527 |
| YALSWVR | YAMSWVR | 3-23 | 3500 | 1 | 0 | 3.75E+10 | 3.75E+10 | 2.24E+11 | 0.00032 | 0.01442 | 0.01442 |
| YALSWVR | YAMSWVR | 3-47 | 3100 | 1 | 0 | 2.45E+10 | 2.45E+10 | 2.45E+10 | 0.00020 | 0.00943 | 0.00943 |
| YALSWVR | YAMHWVR | 1-3 | 3000 | 2 | 0 | 3.23E+08 | 3.23E+08 | 3.23E+08 | 5.83E-08 | 0.00012 | 0.00012 |
| YALSWVR | YAMHWVR | 3-9 | 3000 | 2 | 0 | 3.23E+08 | 3.23E+08 | 3.23E+08 | 5.83E-08 | 0.00012 | 0.00012 |
| YALSWVR | YAMHWVR | 3-30-3 | 3000 | 2 | 0 | 3.23E+08 | 3.23E+08 | 3.23E+08 | 5.83E-08 | 0.00012 | 0.00012 |
| YALSWVR | YAMHWVR | 3-43 | 3000 | 2 | 0 | 3.23E+08 | 3.23E+08 | 3.23E+08 | 5.83E-08 | 0.00012 | 0.00012 |
| YALSWVR | YAMSWFR | 3-49 | 3000 | 2 | 0 | 3.23E+08 | 3.23E+08 | 3.23E+08 | 5.83E-08 | 0.00012 | 0.00012 |
| YALSWVR | YAMHWVR | 3-64 | 3000 | 2 | 0 | 3.23E+08 | 3.23E+08 | 3.23E+08 | 5.83E-08 | 0.00012 | 0.00012 |
| YALSWVR | YAMNWVR | 7-4-1 | 3000 | 2 | 0 | 3.23E+08 | 3.23E+08 | 3.23E+08 | 5.83E-08 | 0.00012 | 0.00012 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |
| YYVDSVK | YYVDSVK | 3-7 | 3400 | 0 | 0 | 2.29E+12 | 2.29E+12 | 2.29E+12 | 0.47474 | 0.47473 | 0.47473 |
| YYVDSVK | YYVDSVK | 3-52 | 3400 | 0 | 0 | 2.29E+12 | 2.29E+12 | 2.29E+12 | 0.47474 | 0.47473 | 0.47473 |
| YYVDSVK | HYVDSVK | 3-16 | 3200 | 1 | 0 | 2.74E+10 | 2.74E+10 | 2.74E+10 | 0.00568 | 0.00567 | 0.00568 |
| YYVDSVK | YYADSVK | 3-23 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-30 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-30-3 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-30-5 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-33 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-43 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-53 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-66 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-69-1 | 3000 | 1 | 0 | 2.19E+10 | 2.19E+10 | 2.19E+10 | 0.00453 | 0.00453 | 0.00453 |
| YYVDSVK | YYADSVK | 3-64 | 2900 | 1 | 0 | 1.94E+10 | 1.94E+10 | 1.94E+10 | 0.00402 | 0.00402 | 0.00402 |
| YYVDSVK | HYADSVK | 3-35 | 2800 | 2 | 0 | 2.54E+08 | 2.54E+08 | 2.54E+08 | 5.26E-05 | 5.26E-05 | 5.26E-05 |
| YYVDSVK | YADSVK | 3-20 | 2700 | 2 | 1 | 3.31E+06 | 1.98E+07 | 6.42E+07 | 6.85E-07 | 4.10E-06 | 1.33E-05 |
| YYVDSVK | YYADSV | 3-47 | 2600 | 2 | 1 | 2.90E+06 | 1.73E+07 | 5.62E+07 | 6.00E-07 | 3.59E-06 | 1.17E-05 |

**- *Distance: It’s all relative***

Gah

Proposals

**- 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, is restricted. In particular, blastp seems to set a hard-coded threshold for score output. This would not be an oversight for simple protein identification, but consider the following homology matches -

**- Blastp custom substitution matrix**

The BLOSUM62 substitution matrix 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.



**- Blastp custom conservation weighting**

The BLOSUM62 substitution matrix described above further assumes that the rate of substitution is spatially uniform - That is, any amino acid is equally likely to be substituted than any other. It is known that this is not the case for immunoglobulins, with a relatively greater likelihood of mutation 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 processes**

The algorithmic differences between de novo and database peptide assignment are suspected to confer some level of error to spectral assignment. It is not known what factors are responsible for incorrect assignment in either case.

Production

Version: v0.9.6

Release: 2016-09-17

Codebase: 4,857 source lines of code

Dependency: 18 files

Version history: 139 commits

Codebase additions: 21,040 source lines of code

Codebase deletions: 16,083 source lines of code

Development environment: Microsoft Visual Studio Community 2015

Version control: Git

References

A

A

A

Workshop

“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.”