



Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium  
Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium  
European Molecular Biology Laboratory Outstation, European Bioinformatics Institute,  
Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom  
<http://www.proteomics.be>

## ICELOGO MANUAL

Niklaas Colaert  
Kenny Helsens  
Lennart Martens  
Joël Vandekerckhove  
Kris Gevaert

<http://icelogo.googlecode.com/>

# Contents

<b>Introduction</b>	<b>iii</b>
<b>I Part one, the installation.</b>	<b>1</b>
<b>1 General Information</b>	<b>2</b>
1.1 Downloading iceLogo . . . . .	2
1.2 Installing and running iceLogo . . . . .	2
1.2.1 Installing iceLogo . . . . .	2
1.2.2 Running iceLogo . . . . .	2
<b>2 Java</b>	<b>3</b>
2.1 Previous Java installation? . . . . .	3
2.2 New Java installation . . . . .	3
<b>II iceLogo</b>	<b>5</b>
<b>3 Running iceLogo</b>	<b>6</b>
3.1 Static reference method . . . . .	7
3.1.1 Positive set . . . . .	8
3.1.2 Reference set . . . . .	8
3.1.2.1 Fixed reference set . . . . .	9
3.1.2.2 Proteome background reference set . . . . .	9
3.1.3 Statistics used . . . . .	10
3.2 Sampling . . . . .	10
3.2.1 Wizard Step 1 - The Reference set . . . . .	11
3.2.1.1 Statistics . . . . .	12
3.2.1.2 Sampling Types . . . . .	12
3.2.1.3 User input . . . . .	14

---

3.2.2	Wizard Step 2 - The Experimental set . . . . .	15
3.2.2.1	User input . . . . .	16
3.2.3	Wizard Step 3 - The Overview . . . . .	18
<b>4</b>	<b>Visualization methods in iceLogo</b>	<b>21</b>
4.1	iceLogo . . . . .	21
4.2	Bar Chart . . . . .	22
4.3	Heat map . . . . .	23
4.4	Sequence logo . . . . .	24
4.5	subLogo . . . . .	26
4.6	Amino acid parameter graph . . . . .	27
4.7	Correlation line . . . . .	28
<b>5</b>	<b>Parameters</b>	<b>30</b>
5.1	Choosing visualization output . . . . .	30
5.2	P-value . . . . .	30
5.3	Amino acid color . . . . .	31
5.4	General parameters . . . . .	31
5.5	Aa parameter chooser . . . . .	34
5.6	Saving the output visualization panels . . . . .	35
5.7	Species adder . . . . .	35
<b>6</b>	<b>Problems and questions</b>	<b>37</b>

# Introduction

## iceLogo

Knowing and understanding an amino acid consensus sequence of a functional part of a protein is important in biochemistry. Sequence logos are, until today, the most used tools to visualize consensus sequences. These use multiple sequence alignments as input and are based on *Shannon's information theory*. A web based application (<http://weblogo.berkeley.edu/>) was developed to generate sequence logos in a fast and easy way. A sequence logo is a histogram-like figure where every bar is replaced with a vertical stack of letters. The total height of the stack is expressed in *bits* and the height of one letter in that stack is proportional to its frequency at that specific position. The maximal height of a stack is 4.32 *bits* when using amino acid sequences and equivalent to the occurrence of only one amino acid on a position. The higher the stack, the more conserved this position is in the input multiple sequence alignment.

Considering the disadvantage of sequence logos is that the probability of occurrence for every amino acid is the same, 0.05 (5%). This disadvantage is partially resolved by structure logos. In structure logos, a calculation is performed to correct for the user-defined amino acid frequencies. However, structure logos still use the information theory and the outcome is still expressed in the less informative *bits*.

Comparisons of two data sets can be done by the Two Sample Logo web application (<http://www.twosamplelogo.org/>). This tool lets the user find under- or over-represented amino acids in one of two sets when compared to each other. Statistical tests are used in Two Sample Logos to reduce the number of displayed amino acids by only presenting statistically significant amino acids.

The previous tools let the user find patterns and consensus sequences in the input set. Still, there are two major drawbacks with these methods. First, these tools were designed for finding patterns in DNA or RNA. In proteins, the difference between high and low abundant amino acid is much higher than in nucleic acids. These tools cannot give a clear view on the highly

important low abundant amino acids. A method to not only show absolute but also relative differences (ex. fold change) is also needed. Second, no tools exist where significant over- and under-represented amino acids can be found when compared to a proteome background. Here, we describe a program, iceLogo, that counters these shortcomings

## **About this manual**

The first part of this manual describes the installation of Java and iceLogo. The second part explains how iceLogo can be used.

## Part I

### Part one, the installation.

# Chapter 1

## General Information

### 1.1 Downloading iceLogo

The main iceLogo site is <http://icelogo.googlecode.com/>. There, under the download section, an installer can be downloaded. The source code can be acquired via subversion.

### 1.2 Installing and running iceLogo

#### 1.2.1 Installing iceLogo

Click the iceLogo-1.0-installer.jar and let the install wizard do the rest.

#### 1.2.2 Running iceLogo

For Microsoft Windows systems: in the installation directory choose the iceLogo.exe file and double click it. For Unix-based systems: in the installation directory, start the iceLogo.sh file. If iceLogo does not start, try double clicking the iceLogo.jar file in the core/lib folder.

The virtual memory that is committed to the iceLogo program can be changed by right clicking on icelogo.bat or icelogo.sh and choosing edit. The default value is 512 Mb. This can be changed by editing the variable after *-Xmx* in the last line of the file.

# Chapter 2

## Java

Since iceLogo was developed in Java, it runs on every operating system that has a Java Runtime Environment equal or higher than version 1.5.

### 2.1 Previous Java installation?

It might be that a functional Java is already installed on your computer due to the widespread use of Java nowadays.

If you do not know if you have Java 1.5 or higher installed, you can do the following simple check:

- Open a command window by  
`start` → `run` and enter `cmd`.
- Enter `Java -version` in the command window.

If Java is already installed, you will see something like below where x stands for the version.

```
{Java version "1.x.0_01"}
```

If your Java version is lower than 1.5 you have to upgrade Java. If Java is not installed, you have to make a new installation. In both cases, you have to install a new Java version as explained in 2.2.

### 2.2 New Java installation

If Java 1.5 or higher is already installed, you may skip this step. The installation of Java is quite simple.

- Go to <http://java.com>



- follow the main download link and start the download
- when finished, open the installer and follow the straightforward instructions

Everything you need from Java should be properly set now. Proceed to the next step.

# Part II

## iceLogo

## Chapter 3

# Running iceLogo

In general, iceLogo will use a reference set to calculate the chance of occurrence (p-value) of every amino acid on every position in the experimental set. Choosing a relevant reference set is therefore of great importance and should be tailored to the expected technical and biological background. Basically, there are three ways to create a reference set and these are discussed below and main points of attention are indicated.

Firstly, a static reference set holds the frequency of each amino acid using typically species-specific sequence data extracted from the UniProt/SWISS-PROT database. One may consider using such a dataset when for instance shotgun (thus non-targeted) proteomics data is interrogated for consensus sequence patterns or consensus patterns of amino acid biophysical and biochemical properties. It should nevertheless be noted that such general proteome reference sets may not reflect the expected biological or technical background. Hence, this issue is ought to be diminished by the other methods for reference set creation.

Secondly, a set of user-defined peptide sequences can also be used to construct a multiple sequence alignment as the reference set. For instance, upon analyzing data from a phosphoproteomics experiment, the non-phosphorylated peptides might make up a good reference set to analyze against the phosphorylated peptides and thereby reducing protocol-related bias. Note that the size of this reference set may change the statistical calculations (see 3.1.3 for more information on sample size).

Thirdly, the reference set can be sampled on-the-fly from a FASTA-formatted protein sequence database and this method has both its pros and its cons. The main disadvantage is a time cost since the FASTA file is repeatedly iterated while sampling and this is an intense computational process. The main advantage however counteracts the issue that upon using static amino acid frequencies one assumes that amino acid usage in the experimental dataset is generally equal to


that of the whole proteome. This assumption is taken for granted when using a static reference set but might be prone to errors since for instance the amino acid frequencies at the N-terminal part of proteins differ from the globally averaged amino acid frequencies. This type of variation in amino acid usage is neutralized by the sampling method. Another advantage is that online tools exist that allow creating specific FASTA files and thus reference sets that are particularly specific and highly tailored to the expected technical or biological background.

Examples of such tools are:

- The Sequence Retrieval System (SRS) on the EBI site (<http://srs.ebi.ac.uk/>) which is very useful for the creation of highly specific FASTA-formatted protein database from all major protein databases.
- Similar to SRS, the query builder tool on the UniProt website (<http://www.uniprot.org>) allows the user to create SRS-like and Google-like (simpler) queries to create specific FASTA files.
- Database on Demand (<http://www.ebi.ac.uk/pride/dod/>) is a web-based database pre-processing tool that will generate custom FASTA formatted sequence databases according to a set of user-selectable criteria using the commonly used UniProt/SWISS-PROT and UniProt/TrEMBL databases as sources.

A last critical note on iceLogos feature of customized reference set definitions is that although this feature enables an user to translate her/his hypothesis into a reference sequence set, against which she/he can test an experimental sequence set, several things may go wrong when creating this reference set. Failing criticism upon creating the reference set might lead to wrong conclusions. As a rule of thumb, remind yourself that every constraint or rule used to create the reference set directly influences the conclusions to be drawn about the experimental sequence set.

### 3.1 Static reference method

This method can be chosen by clicking the *static* button . The static method panel will be visualized in the iceLogo frame and can be seen in figure 3.1.

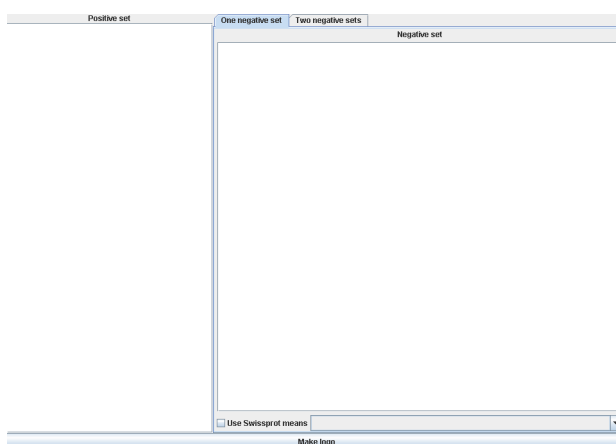


Figure 3.1: The static input panel.

### 3.1.1 Positive set

The positive set in the static method always a multiple sequence alignment. The region of interest must always be located on the same position in every line and the different sequences in the multiple sequence alignment must always be of the same length. Gaps in the alignment will be replaced by X's. This set can be entered in the left text field on the input panel (see figure 3.1)

### 3.1.2 Reference set

Two different ways can be chosen to create a static reference set and are explained in section 3.1.2.1 and section 3.1.2.2. Also, not one but two reference sets can be used. Two reference sets can be useful if only at certain positions a certain technical bias is applicable. One for the first positions and one for the last positions in the positive set. The use of two negative sets can be initiated by clicking the *two negative sets* tab at the right side of the frame. Then, a slider and two panels where the negative set can be defined, will be visualised. With the slider, the position where the second reference set starts and the position where the positive set will be split can be defined (see figure 3.2).

One negative set   Two negative sets

Negative set

Split sequences on position: 15

1   6   11   16   21   26

Example: GHISVKEPTPSIASD (length: 15)

Example: ISLPATQELRQRLR (length: 15)

SDELRRQDKSSGASSE  
 SDELRRQKFLEGFADF  
 SDELRRQLAARLEAL  
 SDELRRQVYQRSTASH  
 SDELRRHMRVHTRYR  
 SDELRRRLQREIHKLQ  
 SDELRRVLLVEDSEK  
 SDELRTLSEMERGAQQ  
 SDGALLLGASSLSGR  
 SDKKKSEGGIEIVKE  
 SDPNAVAPAPQGVRL  
 SDQGGKAHSXXXXXX  
 SDSGSSSEPFDRHA  
 SDTDVSMPLVEERHR  
 SDVDLYQVRTARHNI  
 SDVLETVVLINPSDE  
 SEAKDGINRTALREI  
 SEAVVEYVFGSRLK  
 SEDFGVNEADSDA  
 SEDPDQYLLINTAR  
 SEDSTIYDLLKDPVS  
 SEGFDTYRCDRLAM  
 SEIDLFINRKEFRKM  
 SELELTGKLEQVRS  
 SELNLRRLFDANKDR  
 SENESLNQESKRAVE  
 SEPTQALELTEDDIK  
 SEQDEVLVSSSRV  
 SEQSSMSIEFGQESP  
 SESLKDVLRLLPYW  
 SEVDMLKIRSEFKRK  
 SEVELAAALSDKRGL

☒ Use swissprot mean   ☐ Use swissprot mean

Homo sapiens

Figure 3.2: The panel where a double negative set can be created. The first 15 amino acids of the positive set will be compared with the *Homo sapiens* Swiss-Prot natural occurrence. The last 15 amino acids of the positive set will be compared with the reference set given in the right text field.

### 3.1.2.1 Fixed reference set

A user can give a multiple sequence alignment as a reference set. This negative set can be set in the large text field in the *negative set* tab.

### 3.1.2.2 Proteome background reference set

Predefined negative sets are stored in iceLogo. For every species in the Swiss-Prot protein database the proteins will be collected and the frequency of every amino acid will be calculated. Adding a species can be done with the species adder (see section 5.7). The proteome background can be used by clicking the *use Swiss-Prot mean* checkbox and by defining the species to use in the drop-down box next to or below the checkbox.

### 3.1.3 Statistics used

Different parameters must be calculated before iceLogo can decide if the presence of a specific amino acid at a specific position is significant.

**Sample size** The sample size is of great importance in the calculation of the standard deviation.


The sample size is defined by the positive set and the reference set. When the sets are both multiple sequence alignments, the smallest set size (the multiple sequence alignment with the least sequence lines) will be used as the sample size. If the reference set is created with the proteome background method, the size of the positive set will be used as the sample size.

**Standard deviation** The standard deviation ( $\sigma$ ) uses the sample size ( $N$ ) and the frequency ( $f\%$ ) of an amino acid in the reference set and is calculated with formula 3.1.

$$\sigma = \sqrt{\frac{f\%}{N}} \quad (3.1)$$

This calculated standard deviation will be used to calculate significances in the different visualization methods.

## 3.2 Sampling

The second method, creating the reference set from a FASTA file, can be chosen by clicking on . Applying the sampling method requires more parameters than the static method. As such, iceLogo guides the user through these steps in a convenient 3-step wizard.

**Step 1** assists the creation of the reference set and the user is therefore required to define which sampling strategy should be used and which FASTA file reflects the protein sequence background.

**Step 2** assists the creation of the experimental set(s) with an optional anchor position to align the sequence onto the FASTA file.

**Step 3** overviews the configuration by simulating the upcoming analysis.

### 3.2.1 Wizard Step 1 - The Reference set

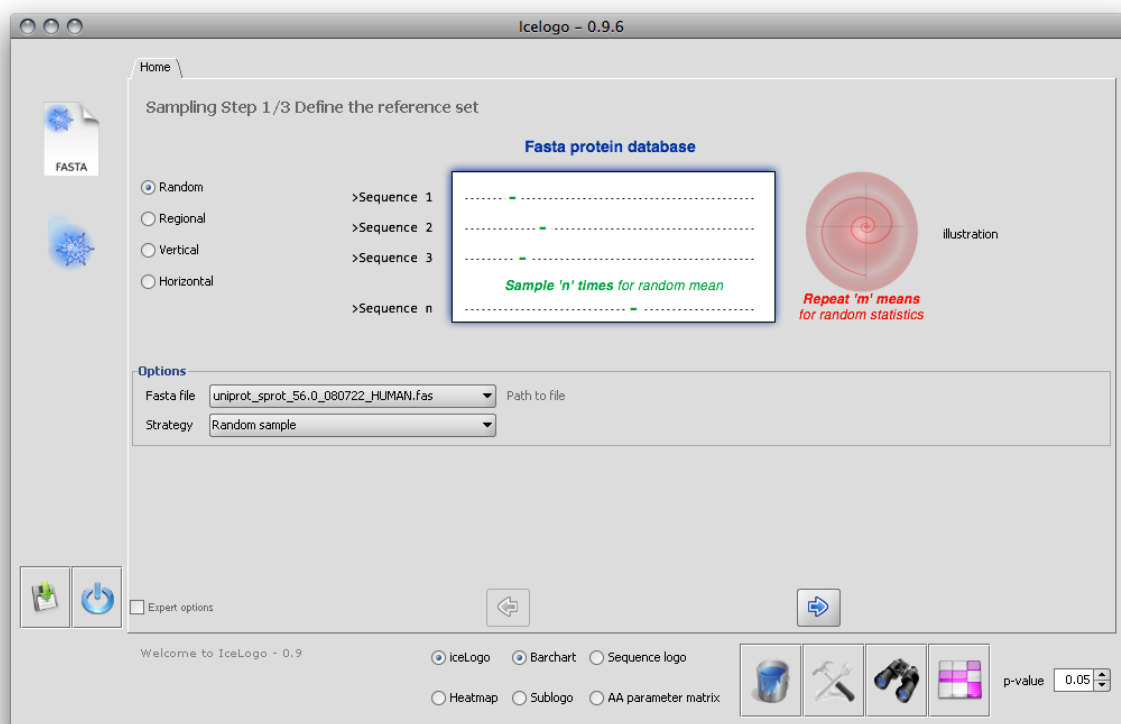


Figure 3.3: This figure illustrates the first wizard step of the sampling analysis in iceLogo

Sampling from a FASTA file comes both with advantages and drawbacks. When using static amino acid frequencies, one assumes that amino acid usage is generally equal to that of the whole proteome. But since this assumption is not actively tested by static methods, these methods might be prone to error. Therefore the major advantage of sampling from a FASTA sequence database is such that (unexpected) variation in amino acid usage is included in the sampling test. For example, Crooks et al. published in 2004 on deviating amino acid usages at protein extremities. Or else, amino acids have an elevated probability to repeat themselves, which might lead an increased translation efficiency. Both examples could lead to unwanted statistical interaction and invalid conclusions.

The major drawback is that the FASTA file must be repeatedly accessed and this computation comes with a time cost. Compared to instantaneously creating a static reference set, the sampling from the human subset of Swiss-Prot might last a minute or more.



Since we reckoned that the retrieval of a FASTA sequence database might be difficult for a layperson, iceLogo also comes with an easy-to-use FASTA retrieval application.

### 3.2.1.1 Statistics

The reference set is the backbone for the statistics by reflecting the probability of finding an amino acid (AA) at random or under certain conditions. This is done as following. If the experimental set contains  $n$  peptides, iceLogo samples  $n$  peptides from a FASTA file and thereby calculates individual amino acid frequencies. If this process is iterated for at least 30 times, then the central limit theorem tells us we are allowed to infer normally distributed reference statistics with a mean and a standard deviation for each amino acid (3.2). Finally the experimental sequence set, also containing  $n$  peptides, can then be tested against this reference distribution and conclusions can be drawn in terms of probability (3.3) by performing a  $t$ -test.

$$N(\mu_{AA}, \sigma_{AA}) \quad (3.2)$$

$$P(AA) = \frac{1}{\sigma_{AA}\sqrt{2\pi}} e^{-(x-\mu_{AA})^2/2\sigma_{AA}^2} \quad (3.3)$$

### 3.2.1.2 Sampling Types

iceLogo has various algorithms to sample peptides from the FASTA file. These are the so called sampling types which the user can choose and are listed below.

Among the distinct algorithms, the following variables are common:

**sample size**  $n$  equals the number of peptides to calculate a single  $Freq_{AA}$  per amino acid.

**iteration size**  $i$  equals the number of times the former calculation is iterated to estimate the mean  $\bar{\mu}_{AA}$  and the standard deviation  $\bar{\sigma}_{AA}$  on the frequency per amino acid.

**Random** The random sampling method calculates the probability to encounter an amino acid at random in the FASTA file.

To do this, the algorithm reads  $n$  protein sequences at random from the FASTA file. In each protein sequence, one amino acid is chosen at random and added to an amino acid counter. When  $n$  amino acids have been added to this counter the  $Freq_{AA}$  per amino acid is calculated. This process is then iterated  $i$  times to estimate  $\bar{\mu}_{Random_{AA}}$  and standard deviation  $\bar{\sigma}_{Random_{AA}}$ .

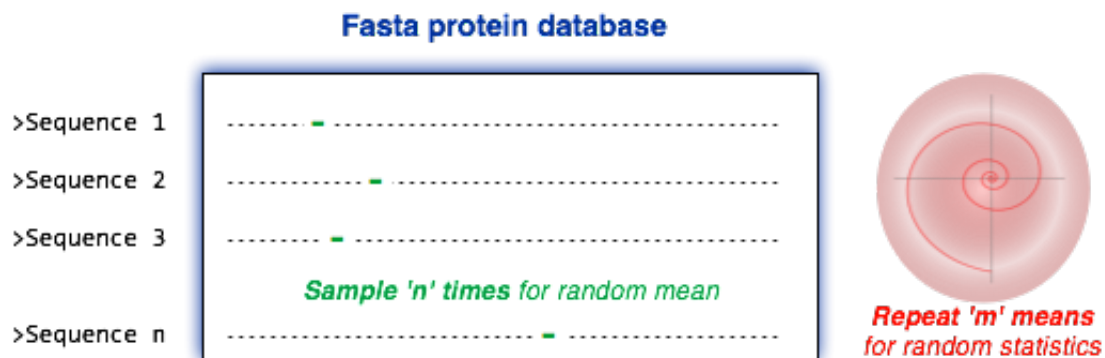


Figure 3.4: This figure illustrates the random sampling algorithm.

**Terminal** The terminal sampling method calculates the probability to encounter an amino acid at a given distance from a protein terminus in the FASTA file.

To do this, the algorithm reads  $n$  protein sequences at random from the FASTA file. In each protein sequence, a terminal peptide is retrieved (N-term or C-term) with length  $l$  equal to the number of amino acids in an experimental peptide. The amino acids are added to  $l$  separate amino acid counter for each position. When  $n$  terminal peptides and their amino acids have been added to these counters, the  $Freq_{AA_i}$  per amino acid is calculated for each position. This process is then iterated  $i$  times to estimate  $\bar{\mu}_{Terminal_{AA_i}}$  and standard deviation  $\bar{\sigma}_{Terminal_{AA_i}}$ .

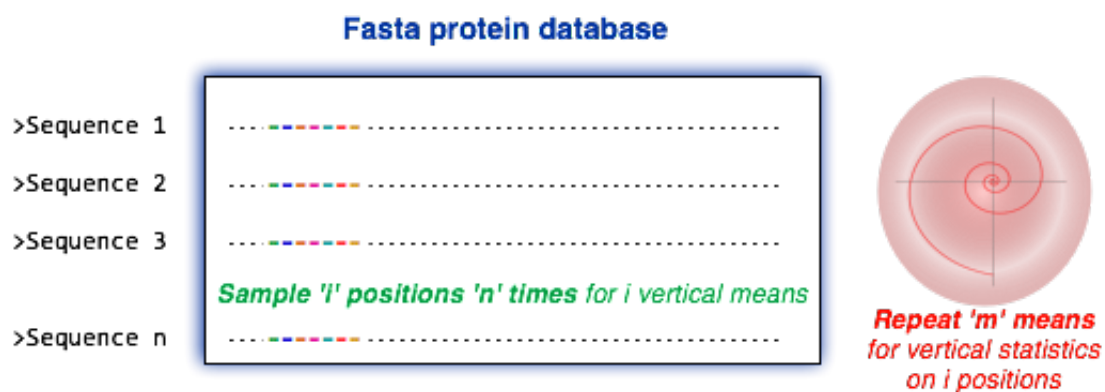


Figure 3.5: This figure illustrates the terminal sampling.

**Regional** The regional sampling method calculates the probability to encounter an amino acid in the region around an anchored experimental position.

To do this, the algorithm first analyses the amino acid frequency at an anchored position  $Freq_{AA_{anchor}}$  in the experimental set. For example, a experimental sequence set with

phosphorylated peptides anchored to the phosphorylation site has  $Freq_{AA_{Ser}} = 70\%$  and  $Freq_{AA_{Thr}} = 30\%$ . Then the algorithm reads  $n$  protein sequences at random from the FASTA file. In each protein sequence, a regional peptide around the anchor site is retrieved with length  $l$  equal to the number of amino acids in an experimental peptide. In the example,  $0.70 \times n$  regional peptides have a Ser anchor and  $0.30 \times n$  regional peptides have a Thr anchor. The amino acids are then added to  $l$  separate amino acid counter for each position around the anchor site. When  $n$  regional peptides and their amino acids have been added to these counters, the  $Freq_{AA_i}$  per amino acid is calculated for each position around the anchor. This process is then iterated  $i$  times to estimate  $\bar{\mu}_{Regional_{AA_i}}$  and standard deviation  $\bar{\sigma}_{Regional_{AA_i}}$ .

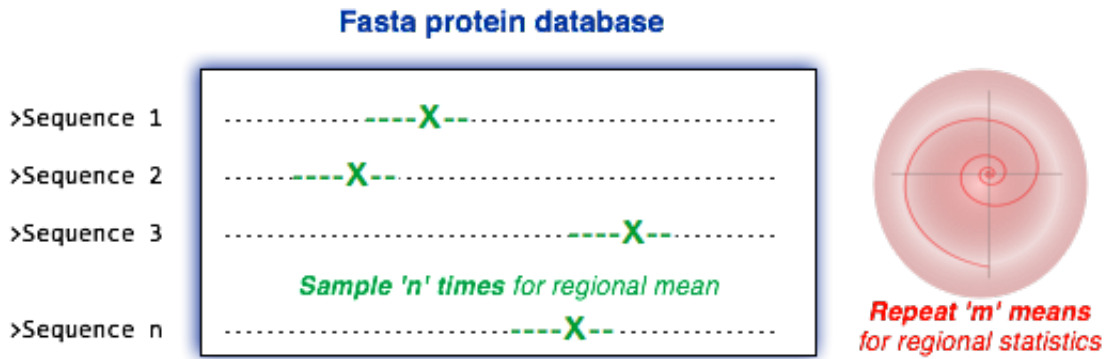


Figure 3.6: This figure illustrates the regional sampling

Note that other Sampling Types can be implemented due to the pluggable software architecture. If required, please post a request in the google user group (<http://groups.google.com/group/iceLogo>).

### 3.2.1.3 User input

Understanding the different sampling types, the creation of a reference set is done by a convenient user interface as shown in 3.3.

**Sample type** On the left side, the sampling type is chosen by clicking the appropriate radio button.

**Illustration** On the right side, the chosen sampling type is illustrated for user convenience.

**Options** On the bottom side, general input options as well as sampling type options are shown.

**Fasta database** The pull-down menu lists the three last used FASTA files. Select the FASTA file that should be used for the sampling analysis. If no FASTA files are listed,

you must add or download a new FASTA file. To *add a new FASTA file*, click the corresponding item and navigate to the file on your computer. To *download a new FASTA file* click the corresponding item and wait for the FASTA downloader dialog to open (3.7). Therein, search the appropriate species and click the *Save* button to start downloading the FASTA file. iceLogo will then download the latest Swiss-Prot release and filter the species-specific proteins into a new FASTA file. This file is then stored in the `../conf/` folder in the iceLogo installation path and is automatically added into the pull-down menu.

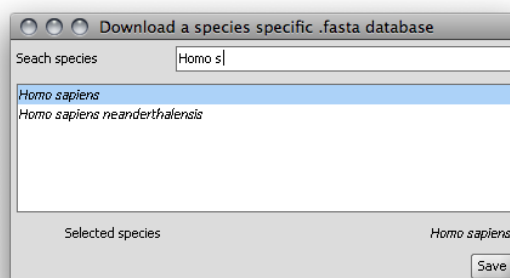


Figure 3.7: This figure shows the FASTA downloader. Start typing a species name (ex: Homo sapiens) in the upper text field and all the matching species names will be matched. By clicking the *Save* button in the bottom-right corner, the appropriate FASTA file will be retrieved from Uniprot.

**Expert Options** These options enable extra fine-tuning of the sampling algorithm but require more understanding of the corresponding algorithm.

**Iteration size** Set the number of iterations for calculating the mean  $\bar{\mu}_{AA}$  and the standard deviation  $\bar{\sigma}_{AA}$  on the frequency per amino acid.

**Terminal - Anchor start position** Set the offset to start terminal sampling

**Terminal - Direction** Set the direction to either sample peptides from the N- or C-terminal end of the protein.

If all parameters have been filled in correctly, click the next button to proceed to the second step which allows you to define the experimental set.

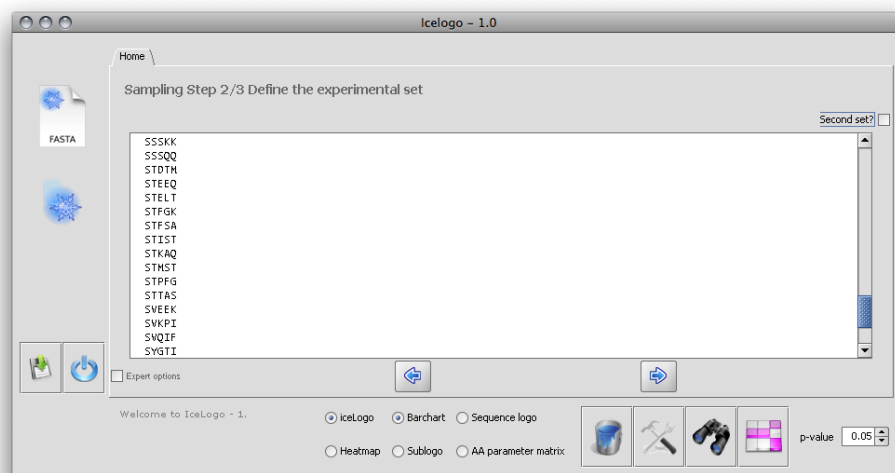
### 3.2.2 Wizard Step 2 - The Experimental set

Opposite to the first step, the second step doesn't require a lot of input: an aligned sequence set that will be tested against the reference set created by sampling a FASTA database. This aligned

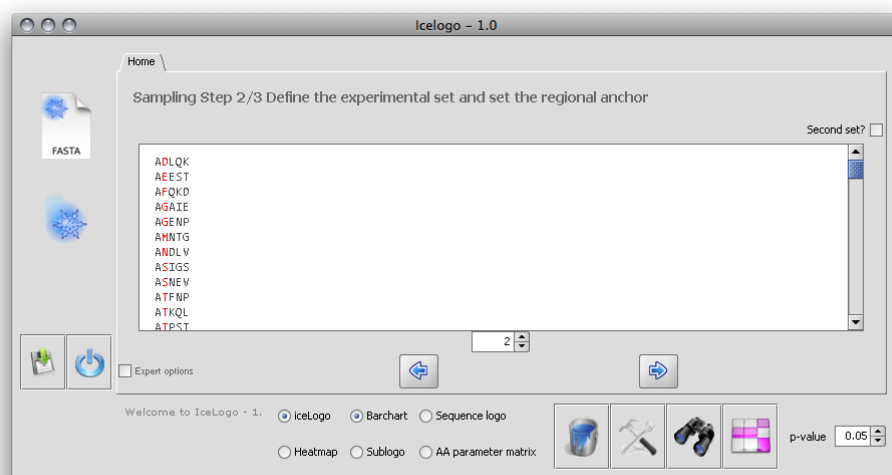
sequence set is typically the result of an experiment (proteolytic cleavage sites, phosphorylation, etc..) and it is therefore called the experimental sequence set.

### 3.2.2.1 User input

**One experimental set** A single experimental (aligned) sequence set can be inserted in the main textfield.



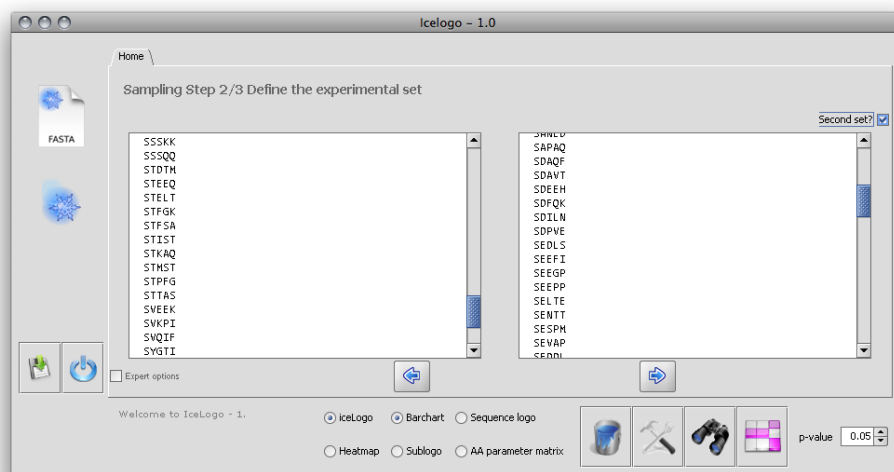
(a) Insert 1 experimental set for *Random* or *Terminal* sampling



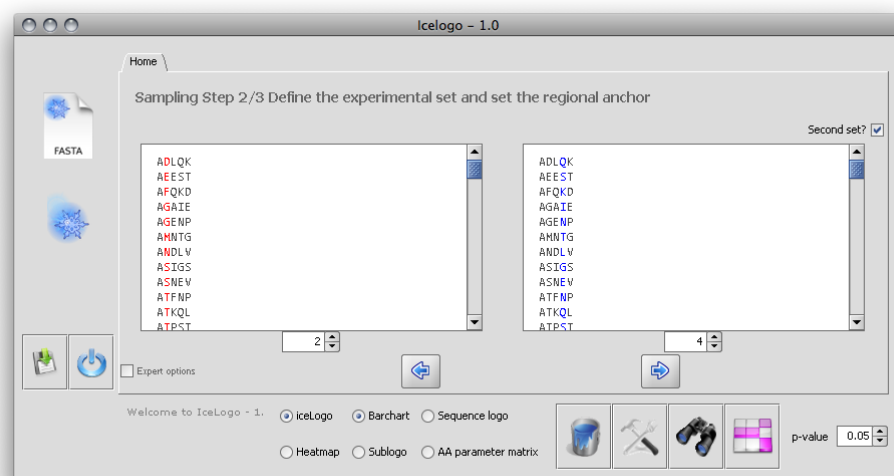
(b) Insert 1 experimental set and its anchor site for *Regional* sampling

Figure 3.8: This figure illustrates the single experimental sequence input dialog. Note that the *Regional* sampling 3.6 also requires the user to define the anchor position.

**Two experimental sets** By clicking the *second experimental set* checkbox in the upper-right corner, a second experimental sequence set can be inserted. Note that not all visualizations are currently support displaying two experimental sets.



(a) Insert 2 experimenal sets for *Random* or *Terminal* sampling



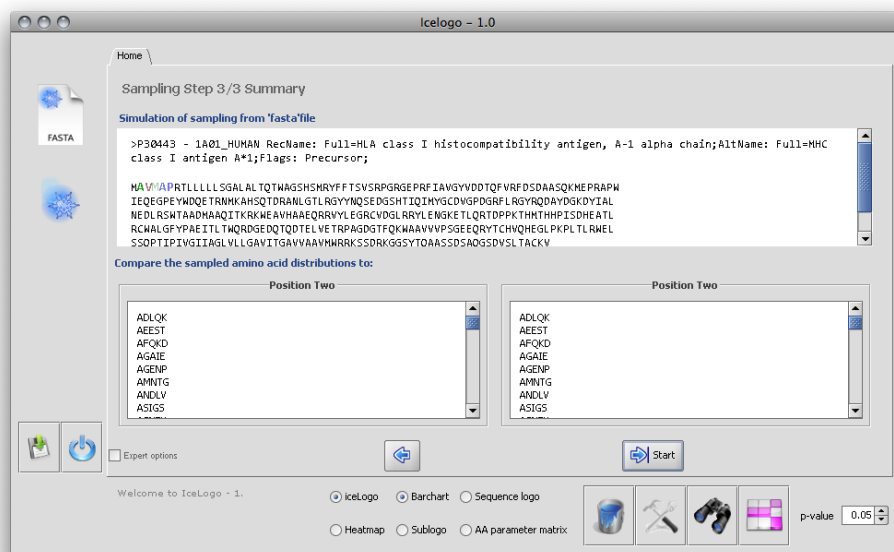
(b) Insert 2 experimenal sets and its anchor site for *Regional* sampling

Figure 3.9: This figure illustrates the double experimental sequence input dialog. Note that the *Regional* sampling (3.6) also requires the user to define the anchor position on both sequence sets.

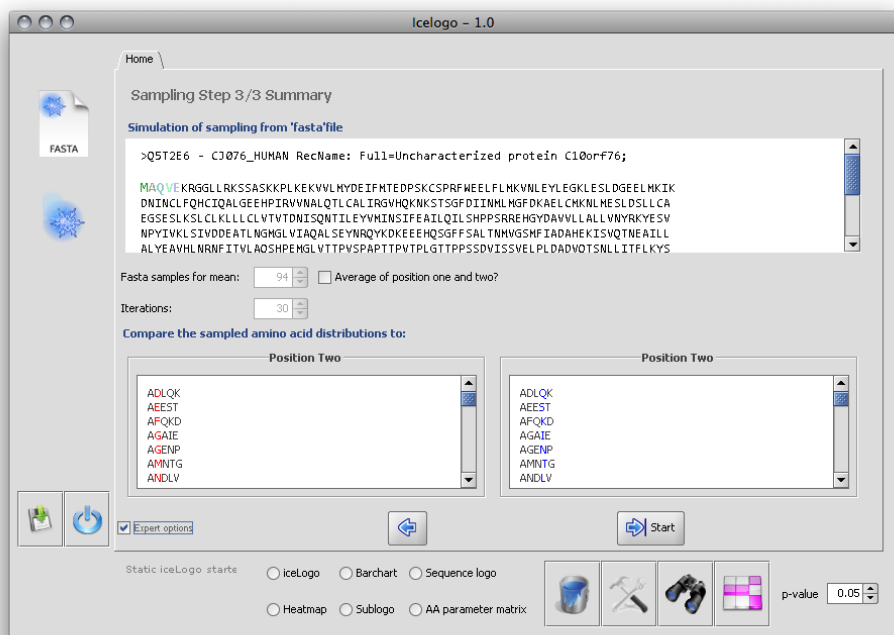
If all parameters have been filled in correctly, click the next button to proceed to the third and last step which presents an overview of the upcoming analysis.

### 3.2.3 Wizard Step 3 - The Overview

In opposite to the former step, the Overview step requires no further input but presents - for user convenience - an aggregation of the user input gathered by the former steps.



(a) The overview of the reference and experimental set.



(b) The overview of the reference and experimental set in expert mode.

Figure 3.10: This figure shows the final Overview step of the sampling Wizard. The upper panel simulates protein entry reading from the FASTA file and meanwhile, color annotates the to be sampled amino acids. Every second, a new protein is displayed to illustrate the sampling process to the user. The bottom panel then shows the experimental sequence sets. In the expert mode, the middle panel also shows sample size  $n$  and iteration size  $i$  (3.2.1.2). Hence both are non-editable since these values are inferred from former input.



**Start!** If the user agrees with the Overview, then the sampling analysis can be started by clicking the next button. Note that this can be, from a computational viewpoint, rather intensive and one should expect the sampling analysis to take one, up to a few minutes.

## Chapter 4

# Visualization methods in iceLogo

The iceLogo program will create a framework (see chapter 3) with a positive and reference set independently of the chosen mode (static or dynamic). By this, iceLogo can use the foundation to create the different visualization methods described below.

### 4.1 iceLogo

An iceLogo attempts to visualize a consensus sequence in a comprehensive manner just like sequence logos. However, it has two major benefits when compared to sequence logos. First, an iceLogo will always use a reference set. When no multiple sequence alignment can be given to create a reference set the option is to use the proteome background or sample a reference set from a protein FASTA file. In this way, iceLogo always uses statistics to find over- and under-presented amino acids. Especially, the visualization of significantly under-represented amino acids is new and is not present in sequence logos. Second, the dynamic nature of iceLogos, mainly the changing of the scoring system (see below), lets the user find changes in low abundant amino acids.

On the iceLogo panel, significantly under- and over-represented amino acids will be visualized. For every position, the amino acid frequencies in the positive set will be compared with the frequencies in the reference set. An amino acid will be regulated if the Z-score is not a part of the confidence interval (for the calculation of the confidence interval see 5.2). The Z-score is calculated with the formula:  $Z\text{-score} = \frac{X - \mu}{\sigma}$ . The formula will calculate how many times the frequency (X) is deviated from the mean ( $\mu$ , the frequency of a specific amino acid on a specific position in the reference set) in terms of the standard deviation ( $\sigma$ ). The way these standard deviations are calculated depends on the iceLogo mode used (see chapter 3).

The color of the amino acids can dynamically be changed in the color panel (see 5.3). The amino acids will be colored pink if the amino acid is significantly regulated, and if this specific amino acid does not occur in the positive or reference set. Three different scoring systems were developed (see 5.4) and can be dynamically changed to generate different iceLogos. If the scoring method is set to *fold change*, the calculated height of a pink amino acid is infinite. Therefore, the height will be set to a specific value. Different scenarios exist for calculating this height.

If only **one** amino acid is regulated and the calculated amino acid size is **infinite**, the height of the amino acid will be the same as the maximal height that can be visualized in the iceLogo.

If **more** amino acids are regulated and **all** the calculated amino acid sizes are **infinite**, the height of the amino acids will be the same as the maximal height that can be visualized in the iceLogo divided by the number of regulated amino acids on that position. All the regulated and infinite amino acids must be either over- or under-represented.

If **more** amino acids are regulated but **not all** the calculated amino acid sizes are **infinite**, the height of the infinite amino acids will be 10 % larger than the largest not infinite amino acid.

The user can zoom in and out on the iceLogo by clicking the left and right mouse button respectively. The height of the Y-axis can also be manually changed in the general parameters panel (see 5.4).

## 4.2 Bar Chart

The bar chart displays absolute frequencies for all amino acids. It is inherent to the iceLogo bar chart to display a single position, therefore the user can control the position by moving the slider at the bottom of the bar chart. The black bars show the mean as calculated from the reference set, along with blue error bars showing the probability boundaries (defined by the p-value in the bottom right corner).

As such, the user is forced to reflect on large standard deviations when using small experimental data sets. Also, if the experimental data set grows, the error bars shrink. Taken together, absolute means for every amino acids and their error are best viewed in the bar chart.

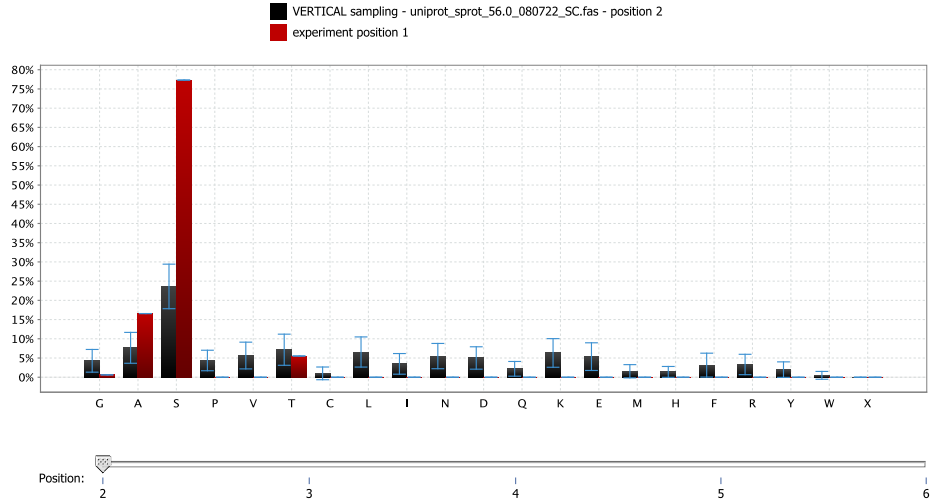


Figure 4.1: The figure shows the bar chart visualisation of an iceLogo analysis. The bar chart is a position specific image and the index of the position can be moved by the slider in the bottom. The top of the bar chart shows a legend of the bars. The black bar represents the amino acid frequency mean and standard deviation within the reference set. The black bars are aligned with colored bars indicating the amino acid frequency within the experimental set.

### 4.3 Heat map

A heat map attempts to visualize all the amino acid occurrences for all positions in one picture. The heat map is a 2D data matrix where every row is an amino acid and every column a position. At the right side of the heat map the gradient shows which p-values correlates with which colour. The Z-score is used for the calculation of the position and amino acid specific p-value and is calculated with the formula:  $Z\text{-score} = \frac{X - \mu}{\sigma}$ . The formula will calculate how many times the frequency (X) of that amino acid on that position is deviated from the mean ( $\mu$ , the frequency of a specific amino acid on a specific position in the reference set) in terms of the calculated standard deviation ( $\sigma$ ). An error function (see formula 4.1) can calculate a p-value for this Z-score.

$$P\text{-value} = \text{erf}\left(\frac{Z\text{-score}}{\sqrt{2}}\right) \quad (4.1)$$

One cell in the heat map matrix will be coloured according to the calculated p-value for that position and amino acid. Only significantly up- and down-regulated elements - according to the given p-value (see 5.2 for setting this p-value) - are coloured in respectively a shade of green and red. The non-regulated elements are coloured black.

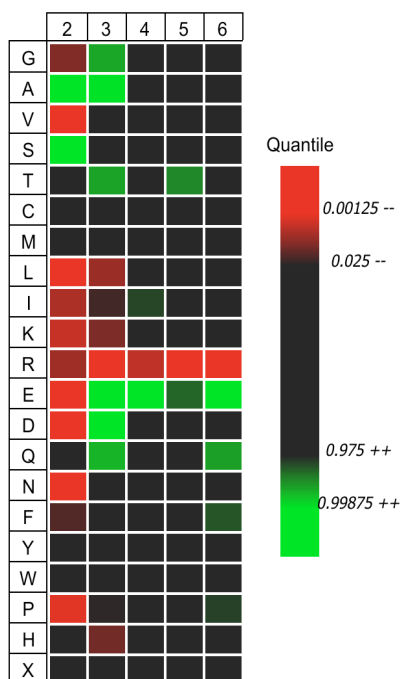


Figure 4.2: The figure shows the heatmap result of an iceLogo analysis. Increased or decreased amino acid frequencies are shown in a gradient of respectively green or red shades.

If the heat map is larger than the displayed screen, it can be visualized by first clicking on the heat map and secondly navigating with the arrow keys.

## 4.4 Sequence logo

Sequence logos were originally created by Schneider and Stephens in 1990 and are used to visualize consensus sequences. Sequence logos are based on the *information theory*. This theory states that a *bit* is the amount of information necessary to choose between two equally probable choices. In a sequence logo the height of a stack of amino acids is thus calculated and presented in *bits*. The height of one amino acid in such a stack reflects its frequency.

The maximal height of the stack is calculate with formula 4.2. Where *choices* stands for the number of possible items. For DNA and RNA this is 4 and thus resulting in a maxBits value of 2. For proteins, there are 20 choices (amino acids) and the resulting maxBits is 4.32.

$$\text{maxBits} = \log_2 \text{choices} \quad (4.2)$$

The final sequence logo height (sH) is calculated with formula 4.3. In formula 4.3 the maxBits is subtracted with the calBits. This calBits is calculated with formula 4.4.  $P_i$  stands for the frequency of amino acid  $i$ .

$$\text{maxBits} - \text{calBits} = \text{sH} \quad (4.3)$$

$$\text{calBits} = -\sum(P_i \log_2 P_i) \quad (4.4)$$

As a simple example, a set of 50 Arg (R) and 50 Lys (K) were used to create a sequence logo. The frequency of these amino acids are in both cases 50%. Formula 4.4 is used to calculate the calBits resulting in 1 (see formula 4.5).

$$\text{calBits} = 1 = -[0.5 \log_2(0.5) + 0.5 \log_2(0.5)] \quad (4.5)$$

The final sequence logo height can be calculated using formula 4.3;  $4.32 - 1 = 3.32$ . The height of the sequence logo in figure 4.3.A is indeed 3.32.

The iceLogo program can use the reference set for a background correction in sequence logos. iceLogo will calculate the height of the stack in the reference set and will subtract it of the height of the stack in the positive set. This corrected sequence logo is presented in figure 4.3.B.

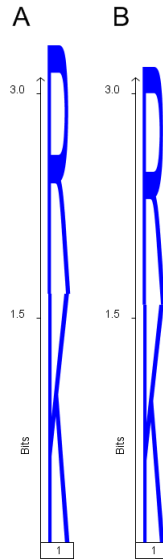


Figure 4.3: These are both sequence logos generated with the iceLogo tool. Figure A is the normal sequence logo. Figure B is the reference set corrected sequence logo. Here the reference set is the human Swiss-Prot proteome.

10000 random human peptides were generated for the following example. The sequence logo without correction is given in figure 4.4.A. Figure 4.4.B gives the sequence logo with reference set correction. With this example, it's clearly shown that a background reduction in sequence logos has an important effect.

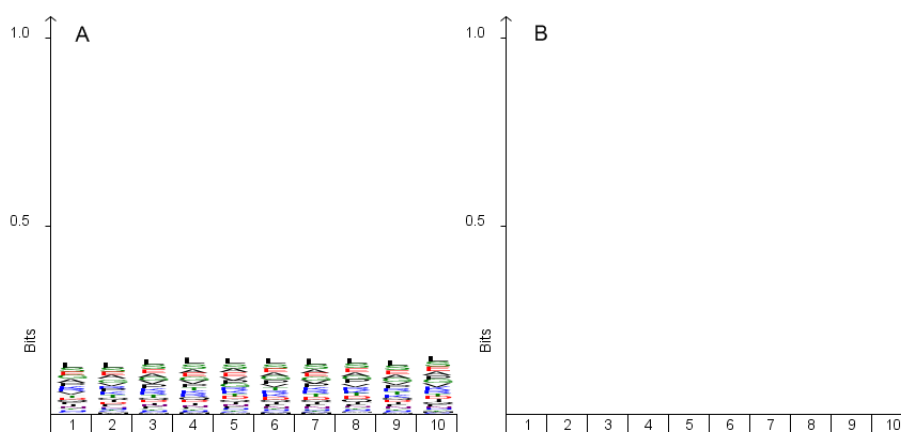


Figure 4.4: These are both sequence logos generated with the iceLogo tool. Figure A is the normal sequence logo. Figure B is the reference set corrected sequence logo. Both were created with 10000 random human peptides.

The user can again zoom in and out by clicking the left and right mouse button. A sequence logo cannot be generated when two experimental sets are compared in the sampling method.

## 4.5 subLogo

In the subLogo tab, the user can try to find hidden elements in the data. The subLogo frame is presented in figure 4.5. In this frame a smaller version of the iceLogo program is run. The user can select a positive set, which is a part of the total positive set. Two ways exist of creating this positive set.

1. The user can select one of the positively regulated amino acids listed in the upper left part of the subLogo tab.
2. The user can select a specific (both regulated and non-regulated) amino acid with the drop-down box and can select a specific position with the position slider both at the left site of the tab.

In both cases, the positive set will be a subselection of the totale positive set where at a specific position, a specific amino acid is present.

The negative set can be created using the totale experimental set or by choosing a Swiss-Prot composition. By default an iceLogo and a bar chart are created. Also, a panel with the subselection of the positive set is created.

The different created panels can be saved by using the *save subLogo* button.

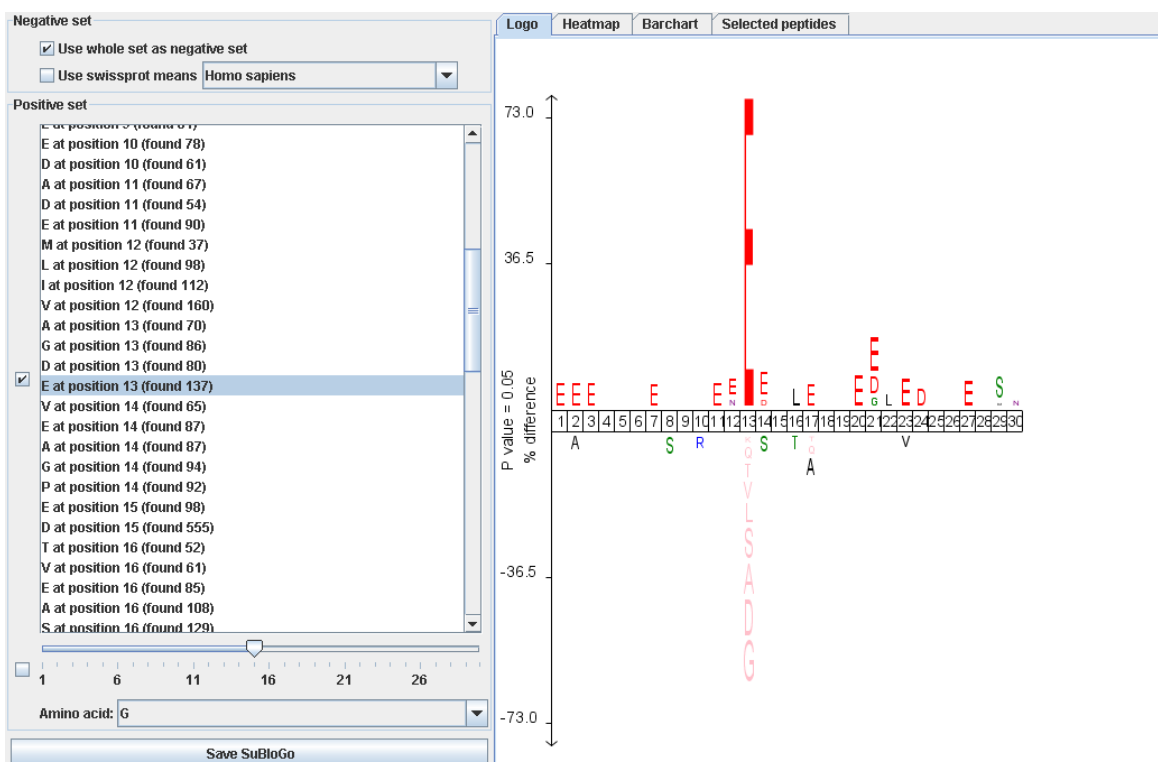


Figure 4.5: The subLogo tab.

## 4.6 Amino acid parameter graph

The iceLogo program can visualize amino acid parameters. These parameters can be found in the AAindex1 database (<http://www.genome.jp/aaindex/>). The AAindex database is a database of numerical indices representing various published physicochemical and biochemical properties of amino acids. Currently, 544 indices are stored as matrices in this database and can be visualised by iceLogo. One of these 544 indices can be chosen in the Aa parameter chooser (see 5.5).

The value for a specific position is calculated with formula 4.6 where  $P_i$  is the frequency for amino acid  $i$  and  $V_i$  is the value for amino acid  $i$  in the amino acid parameter matrix used for the creation of the amino acid parameter graph. The values for different positions from the



experimental set are linked by a green line. If two experimental sets are being analyzed, then this second set is linked by a blue line.

$$\Sigma P_i * V_i \quad (4.6)$$

A sliding window can be used to smoothen the green line that connects the values for the different positions. The minimal size of the sliding window is 2. The sliding window size can be set in the amino acid parameter chooser panel (see 5.5). The value on each position is the mean of the different positions used. In the following table, the positions used for a specific position in the graph are listed for the window sizes 3 and 4.

		Position in graph							
		1	2	3	4	...	28	29	30
		Positions used							
window size	3	<b>1</b>	<b>1,2</b>	<b>1,2,3</b>	<b>2,3,4</b>	...	<b>26,27,28</b>	<b>27,28,29</b>	<b>28,29,30</b>
	4	<b>1,2</b>	<b>1,2,3</b>	<b>1,2,3,4</b>	<b>2,3,4,5</b>	...	<b>26,27,28,29</b>	<b>27,28,29,30</b>	<b>28,29,30</b>

The reference set is used to create a pink zone on the graph. This zone represents the non-regulated region. This non-regulated zone (the confidence interval) is determined by the p-value (see 5.2) and the background standard deviation. Two ways exist to create this background standard deviation.

1. When the static iceLogo method is used (see 3.1) a standard deviation will be calculated for every position. This standard deviation will be calculated on 100 means. One such a mean, is the mean of the amino acid parameter values for N random (based on the reference set) amino acids and N is the sample size (see 3.1.3 for the sample size calculation). This way, the reference set is used to simulate the background for an amino acid parameter.
2. When the sampling iceLogo method is used (see 3.2), a background standard deviation will be calculated for every position. This standard deviation will be calculated on X parameter value means. X is the dimension of sampling. One such a mean, is the mean of the sampled amino acids parameter for a specific dimension. This way, the reference set is used as the background for an amino acid parameter.

The red line in the pink zone represents the mean of the means used for the calculation of the background standard deviation.

## 4.7 Correlation line

The iceLogo program can visualize the correlation between the different amino acids on one position. The correlation is calculated by using a substitution matrix. These substitution matri-

ces can be found in the AAindex2 database (<http://www.genome.jp/aaindex/>). A substitution matrix holds values that describe the rate in which one amino acid changes in another amino acid over time. Currently, 94 substitution matrices are in this database and can be visualised by iceLogo. One of these can be chosen in the general parameter frame (see 5.4).

For every amino acid in the set, the substitution score is calculated by tacking the mean of the substitution values of this amino acids with all the other amino acids in the set on that position. A substitution score is not normalized when the substitution score for one amino acid is multiplied by the substitution value of this amino acid with itself. If the set has 100 amino acids, 99 substitution scores will be calculated for every amino acid. The mean of these 99 substitution score for the different positions from the experimental set are linked by a green line. If two experimental sets are being analyzed, then this second set is linked by a blue line.

The reference set is used to create a gray zone on the graph. This zone represents the non-regulated region. This non-regulated zone (the confidence interval) is determined by the p-value (see 5.2) and the background standard deviation. Two ways exist to create this background standard deviation.

1. When the static iceLogo method is used (see 3.1) a standard deviation will be calculated for every position. This standard deviation will be calculated on 100 substitution means. One such a mean, is the mean of the amino acid substitution values for N random (based on the reference set) amino acids and N is the sample size (see 3.1.3 for the sample size calculation). This way, the reference set is used to simulate the background.
2. When the sampling iceLogo method is used (see 3.2), a background standard deviation will be calculated for every position. This standard deviation will be calculated on X parameter value means. X is the dimension of sampling. One such a mean, is the mean of the sampled amino acids parameter for a specific dimension. This way, the reference set is used as the background for an amino acid substitution.

The dark line in the grey zone represents the mean of the means used for the calculation of the background standard deviation.

# Chapter 5

## Parameters

### 5.1 Choosing visualization output

The different types of output (listed below and discussed in chapter 4) can be selected at the lower end of the frame. The iceLogo and the barchart are selected by default.


- ☒ iceLogo
- ☒ Bar chart
- ☐ Heat map
- ☐ Sequence logo
- ☐ subLogo
- ☐ Aa parameter graph

### 5.2 P-value


The p-value is used in all the different frames and can be set by changing the spinner-box which is located at the right bottom of the screen. The default p-value is 0.05. The confidence interval can be calculated with this p-value using the Wichura algorithm. In the following table some confidence intervals in function of the standard deviation ( $\sigma$ ) are given for p-values.

p-value	< Z-score
...	...
0.20	$[-1.28 \sigma; 1.28 \sigma]$
0.10	$[-1.65 \sigma; 1.65 \sigma]$
0.05	$[-1.96 \sigma; 1.96 \sigma]$
0.02	$[-2.33 \sigma; 2.33 \sigma]$
0.01	$[-2.58 \sigma; 2.58 \sigma]$
0.005	$[-2.81 \sigma; 2.81 \sigma]$
...	...

### 5.3 Amino acid color

The color of the amino acids used in the iceLogo and sequence logo (see sections 4.1 and 4.4) can be changed. This can be done in the panel that opens after the *Amino acid color*  button is pushed.

### 5.4 General parameters

Some general parameters can be set in the panel that opens upon clicking the *general iceLogo parameter*  button. These different parameters will be described in the following and can be seen in figure 5.1.

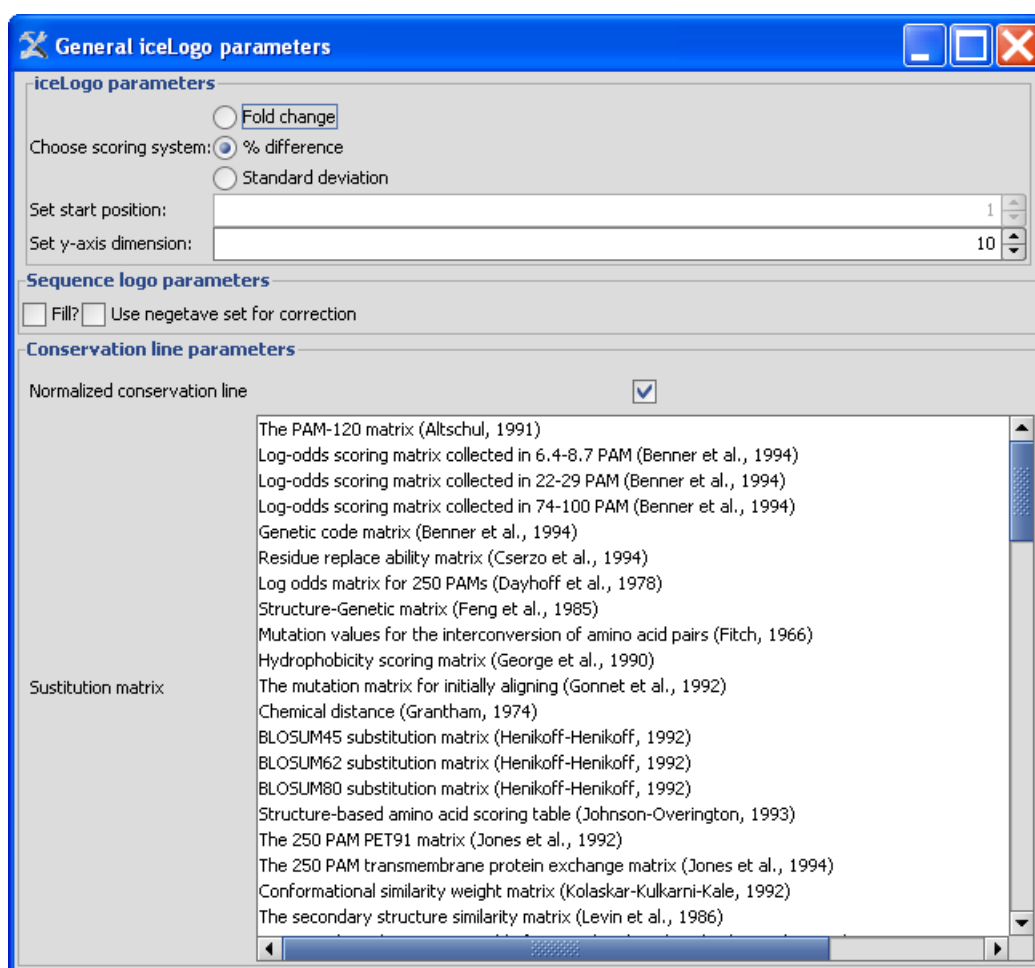


Figure 5.1: The general iceLogo parameter panel.

**Choosing scoring system** Different scoring methods can be used in an iceLogo (see section 4.1). The scoring method has an effect on the size of a regulated amino acid and the vertical position in the stack of regulated amino acids.

- **Fold change** When this method is selected the fold change will determine the size of the amino acid. In the following table the frequencies of two amino acids with their fold change are given. Although the percentage difference between the positive and the reference set is for both amino acid the same (6%), the fold changes of the two amino acid show a large difference ( $7 \iff 2$ ). The fold change scoring method let the user thus look for the regulation of low abundance amino acids.

Type	AA <sub>1</sub>	AA <sub>2</sub>
Frequency in experimental set (F+)	7%	12%
Frequency in reference set (F-)	1%	6%
Percentage difference	6%	6%
Fold change $\frac{F+}{F-}$	<b>7</b>	<b>2</b>

If the calculated fold change (FC) is smaller than 1 the fold change will be converted via formula 5.1 to the converted fold change (FC<sub>con</sub>). By this, the height of negatively regulated amino acid can be compared with the height of positively regulated amino acids.

$$FC_{con} = \frac{1}{FC} * -1 \quad (5.1)$$

The following table gives an example of a converted fold change.

	AA <sub>1</sub>	AA <sub>2</sub>
Frequency experimental set (F+)	12%	6%
Frequency reference set (F-)	6%	12%
Fold change (FC = $\frac{F+}{F-}$ )	2	0.5
Converted fold change (FC <sub>con</sub> )	2	<b>-2</b>

- **Percentage difference** This simple scoring method used the difference in frequency for an amino acid in the experimental set and the reference set as a measure of the height of a letter in the amino acid stack. This is the default scoring method.
- **Standard deviation** The Z-score of an amino acid is used as a measure of the size of a letter in the amino acid stack. The Z-score is calculated with the formula: Z-score =  $\frac{X - \mu}{\sigma}$ . The formula will calculate how many times the frequency (X) is away from the mean ( $\mu$ ) in terms of standard deviation ( $\sigma$ ). The height of the letter will be correlated to the Z-score.

**Start position** The start position used in the different visualization panels can be set with the spinner. The default value is 1. The start position cannot be changed manually if the iceLogo program is run in the sampling mode.

**Y-axis height** The height of the visible part of the Y-axis in the iceLogo (see 4.1) can be set manually with the spinner. Left and right clicking on the iceLogo will also change this value.

**Sequence logo parameters** Two parameters can be used in the creation of the sequence logo.


- **Fill** On the sequence logo panel a filled logo will be created if the *fill* checkbox is selected. A filled logo is a stack of amino acids for every position. The height of

every stack is 100%. The height of every amino acid is related to its frequency. No statistics are used in this panel and all, and not only the significant amino acids, are visualised.

- **Use reference set correction** When this checkbox is selected a sequence logo for the reference set will be calculated. The height of the stack (in *bits*) in the experimental set will be subtracted with the calculated height of the stack in the reference set. See 4.4 for examples.

**Conservation line parameters** In this part of the panel the substitution matrix can be chosen. It can also be selected here that the conservation line should be normalized.

## 5.5 Aa parameter chooser

The amino acids parameter chooser opens after the *chooser*  button is clicked. In the amino acid parameter chooser (see figure 5.2), the list of 544 indices in the AAindex1 database can be searched. This is done by typing a search term in the text field at the top of the panel. If the search term matches a description of an index it will be listed below the search text field. Any of the indices in the list can be selected. The amino acid parameter graph will automatically be updated and extra information like author, pubmed id, amino acid matrix, ... will be shown in the panel.

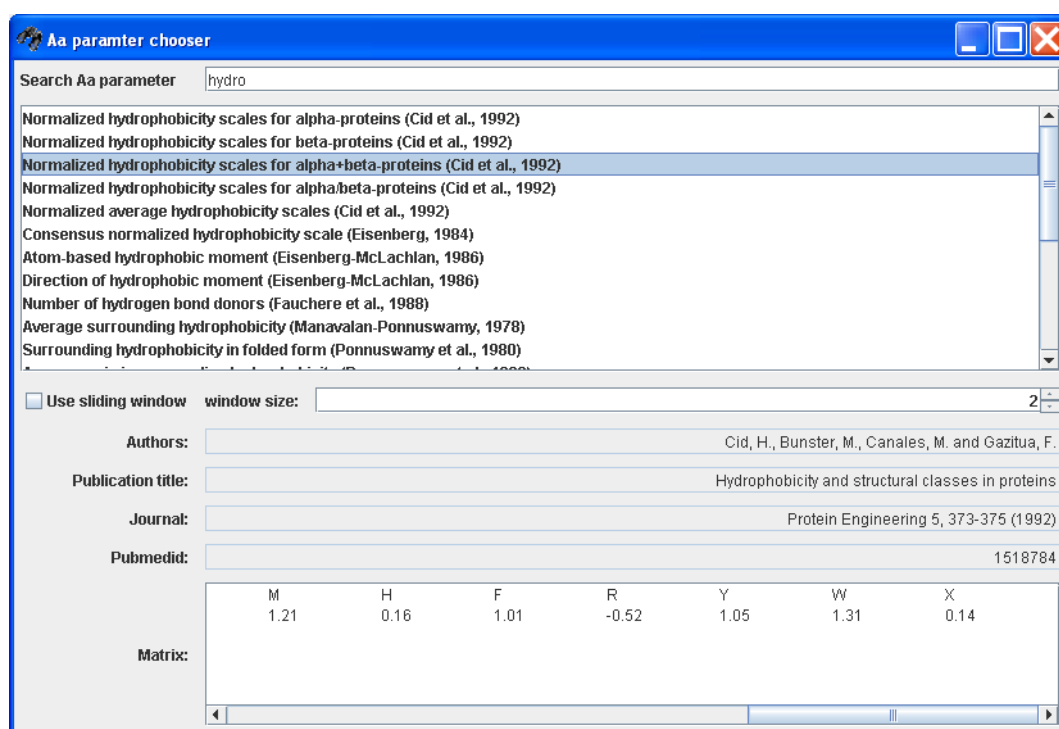




Figure 5.2: The amino acid parameter chooser panel.

The size and the usage of the sliding window can be set in this panel by changing the spinner. The default size is 3.

## 5.6 Saving the output visualization panels

The different panels created with iceLogo can be saved to a *.pdf* file. A save selection panel, where specific visualization panels can be selected, is opened after that the *save* button  is clicked.

## 5.7 Species adder

When iceLogo is run in the static mode (see 3.1) a Swiss-Prot species composition can be chosen to create the reference set. Not all species are listed in the drop-down menu as default. A panel can be opened where other species can be added by clicking the *add species*  button. This panel can be seen in figure 5.3.



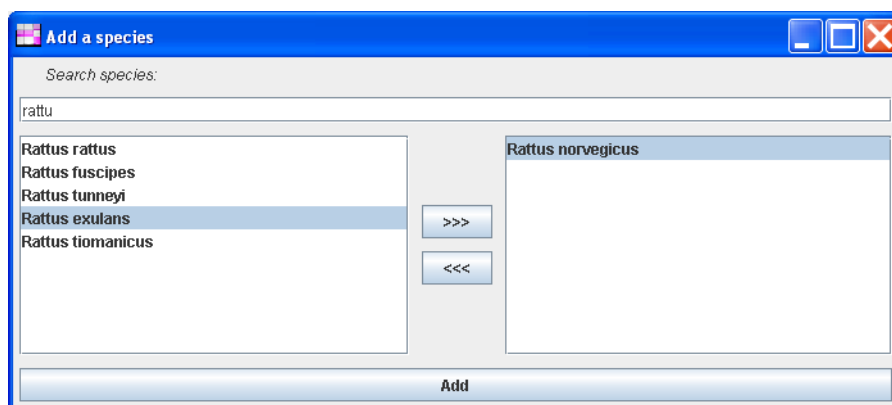


Figure 5.3: New species can be added with this panel.

In the search text field at the top of the panel a part of the scientific name of the species can be typed. The corresponding species will be listed in the left list. Species can be added to the right list by selecting these and clicking the >>> button. By clicking *add* at the bottom of the panel, the species in the right list will be added to the iceLogo program and can be chosen to create the reference set. After adding the new species the program should be restarted.

## Chapter 6

# Problems and questions

A google discussion group (<http://groups.google.com/group/iceLogo>) was created for problems and questions. Also, if you have a request (a new visualization style, ...) you can post a message on the google discussion group.