UMD MANUAL

June 2013 version

Faculté de Médecine Aix*Marseille Université

Inserm
Institut national de la santé et de la recherche médicale

UMD is used by some major international projects in the field of human mutations:

TREAT-NMD Neuromuscular Network

GEN2PHEN

NMDCHIP

THE HUMAN VARIOME PROJECT

HGVS HUMAN GENOME VARIATION SOCIETY

Table of Contents

I- DOWNLOAD OF THE UMD SOFTWARE

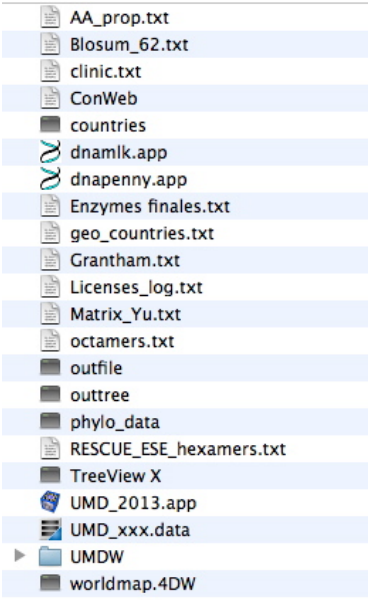
To download the UMD® software, go to the UMD website at http://www.umd.be. On the home page, various information are given:

> It was developed as a generic software to create locus-specific databases (LSDBs) with the 4th Dimension® package from 4D. The UMD software includes an optimized structure to assist and secure data entry and to allow the input of a wide range of clinical data. In addition various analyzing tools have been specifically designed to assist **clinicians** (phenotype-genotype correlations...), **geneticists** (distribution and frequency of mutations...) and **research biologists** (structural domains, molecular epidemiology...). Thanks to the flexible structure of the UMD software, it has been successfully adapted to many genes either involved in cancer (APC, BRCA1, BRCA2, TP53, RB1, MEN1, SUR1, VHL, WT1...) or in genetic diseases (FBN1, LDLR, DMD, VLCAD, MCAD, LMNA, EMD, FKRP, SGCG, SGCA, ATP7B...). This tool is freely available. To download the software please visit the download policy webpage.

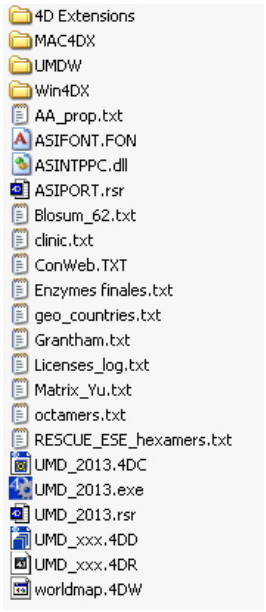
Click on the download hyperlink to access the download policy and register to the UMD software users' database. You will shortly receive an e-mail with a password that will allow you to download the various versions of UMD (Mac/PC) as well as documents.

I-a How to install your UMD software?

Download the file of your choice and copy it to your hard drive: UMD_2013_PC1.zip (PC version) or UMD_2013_MAC.zip (MAC version). Now extract the file (double click on the corresponding zip file). You should get a directory with the followings items:



Mac directory



PC Director

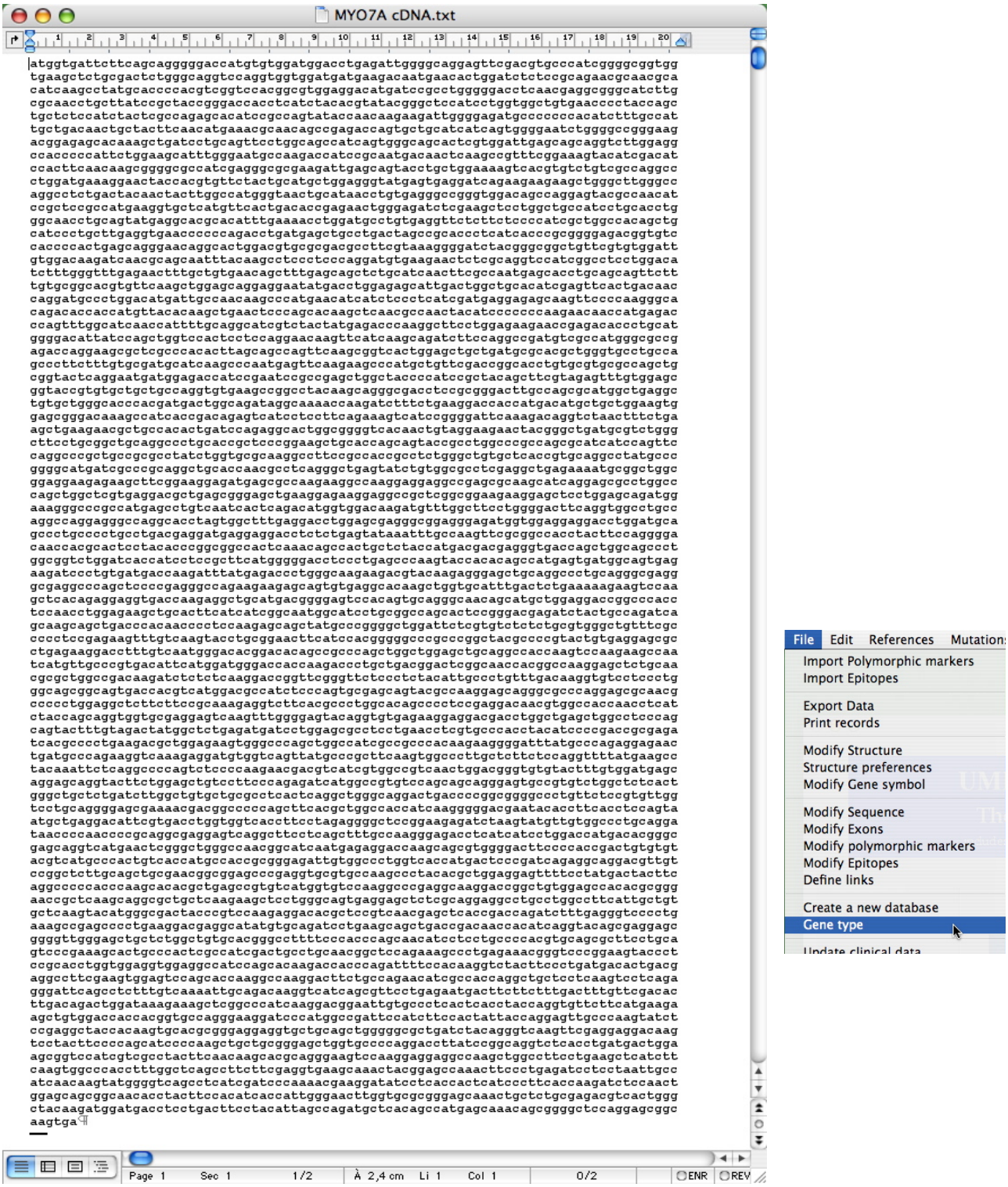
Note that all files and sub-directories should remain in the same directory. In this documentation, all screen shots are from the Mac OSX version of the software but the various steps are identical for the PC version.

I-b How to update your UMD software?

To update the UMD software, download the new UMD® software (see chapter I) and extract the file (see chapter I-a). Copy the data file (UMD_xxx.data for Mac or UMD_xxx.4DD and UMD_xxx.4DR for PC) from your previous version of UMD into this new directory.

II- How to create a new LSDB with UMD?

Before you can start the UMD software, you need to collect additional information such as the cDNA of your gene and its genomic organization. The cDNA should begin by the ATG (translation initiation site) and end by the stop codon.



Example of the MYO7A cDNA. Note that this sequence does not include spaces, numbers or return symbols and was saved as a text file (MYO7A.txt).

You also need to know where are located each exon. For this, you need to define the first nucleotide of each exon using the cDNA reference sequence.

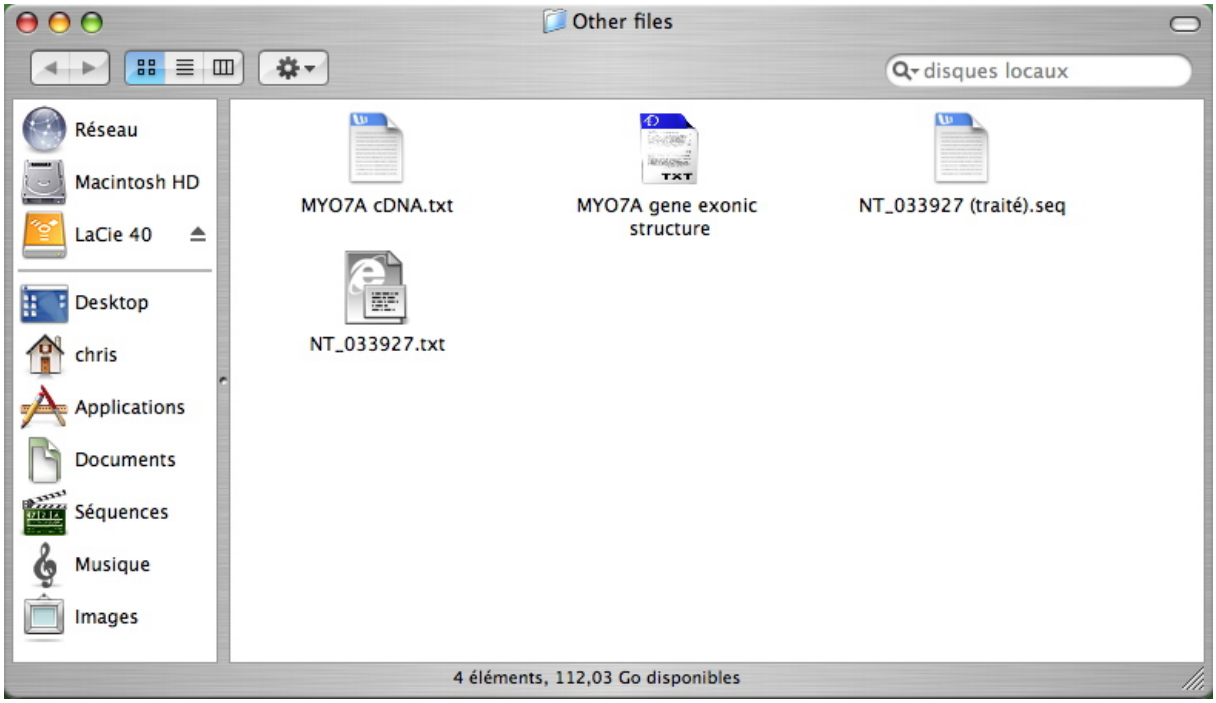
MYO7A gene exonic structure.txt

	A	B	C	D	E
1	1	1	18		
2	2	19	132		
3	3	133	285		
4	4	286	470		
5	5	471	590		
6	6	591	735		
7	7	736	849		
8	8	850	1003		
9	9	1004	1080		
10	10	1081	1200		
11	11	1201	1343		
12	12	1344	1554		
13	13	1555	1690		
14	14	1691	1797		
15	15	1798	1935		
16	16	1936	2094		
17	17	2095	2187		
18	18	2188	2281		
19	19	2282	2366		
20	20	2367	2586		
21	21	2587	2694		
22	22	2695	2904		
23	23	2905	3108		
24	24	3109	3285		
25	25	3286	3375		
26	26	3376	3503		
27	27	3504	3630		
28	28	3631	3750		
29	29	3751	3924		
30	30	3925	4152		
31	31	4153	4323		
32	32	4324	4441		
33	33	4442	4568		
34	34	4569	4852		
35	35	4853	5043		
36	36	5044	5168		
37	37	5169	5326		
38	38	5327	5480		
39	39	5481	5636		
40	40	5637	5742		
41	41	5743	5856		
42	42	5857	5944		
43	43	5945	6051		
44	44	6052	6237		
45	45	6238	6354		
46	46	6355	6438		
47	47	6439	6558		
48	48	6559	6648		
49					
50					
51					
52					
53					

MYO7A gene exonic structure.txt

Exonic structure of the MYO7A cDNA. The position of each exon (first and last nucleotides) is indicated using the cDNA as reference (the A nucleotide of the ATG codon being the first nucleotide).

With the almost completion of the human genome project, genomic contigs have been assembled and intronic sequences of most genes are available. To import these intronic sequences, you can download from the NCBI web site the full contig sequence. You also can download information about SNPs. To do that, please login to the Locus link (http://www.ncbi.nlm.nih.gov/LocusLink/) server and search for your gene. Search for the NCBI Reference Sequences (RefSeq) section and the Category: NCBI Genome Annotation. Follow the link to the Genomic Contig: (for example NT_033927). Save the web page as a text file (using the "save as" menu from your browser). Choose the display FASTA option to get the contig sequence. Save now this sequence as a txt file (use the "save as" menu from your browser). Finally, you need to remove all comments and return symbols from this file.

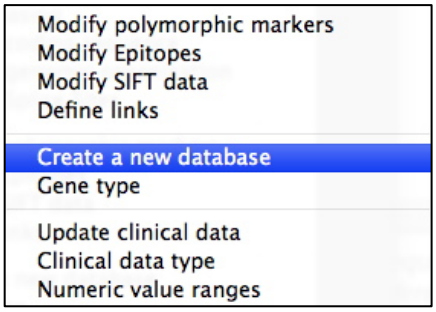


The various reference sequences necessary to build a new LSDB using the UMD software. ***MYO7A cDNA.txt*** *includes the reference cDNA sequence as a text file;* ***MYO7A gene exonic structure*** *the exonic organization of the MYO7A gene according to the reference cDNA;* ***NT_033927.txt*** *is a text file which includes annotations of the genomic contig and* ***NT_033927 (traité).seq*** *is the sequence of the genomic contig.*

You are now ready to build your LSDB using the UMD software.

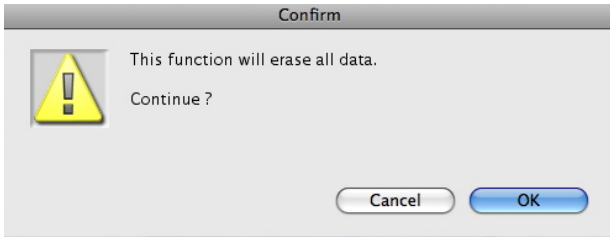
II- CREATE A NEW UMD-LSDB

II-a Set-up your own mutation database using the UMD® software?

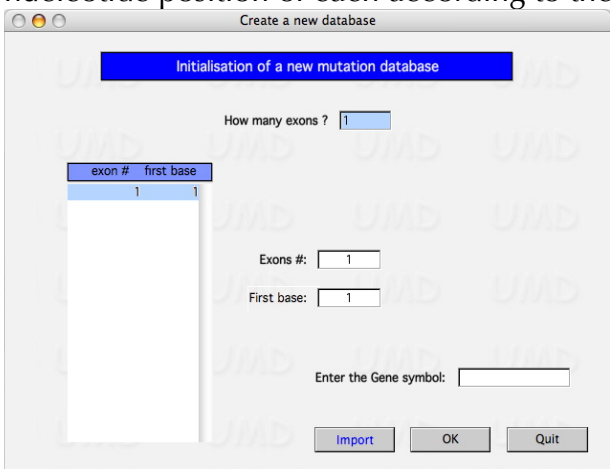


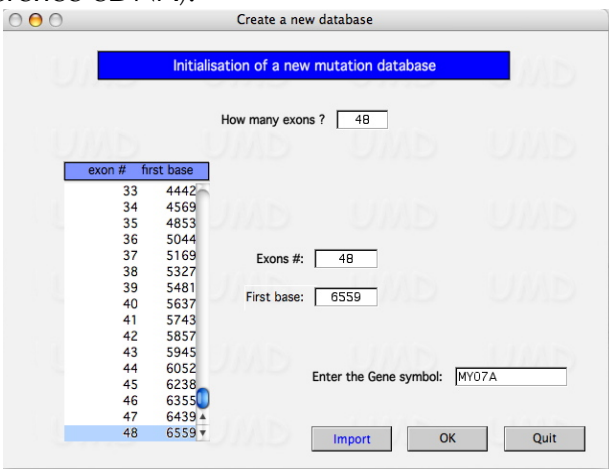
Beware that you can only build a new UMD database from an existing one. Indeed the data file is not empty but contains the genetic code as well as other useful data that can not be directly modified by the user.
Therefore before the creation of a new database, you need to copy the directory of an existing UMD-LSDB and rename it for your new gene. When the new directory is ready, launch the UMD® software and choose the "Create a new database" option from the "File" menu.

As this option will erase all data related to a specific gene (mutations, reference sequences as well as clinical data), an alert message will ask you to confirm this action.

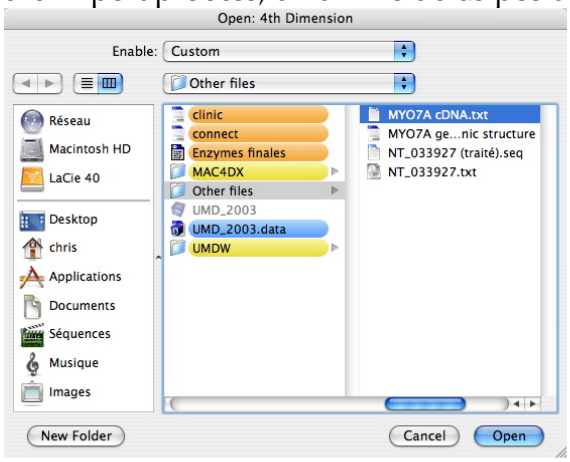


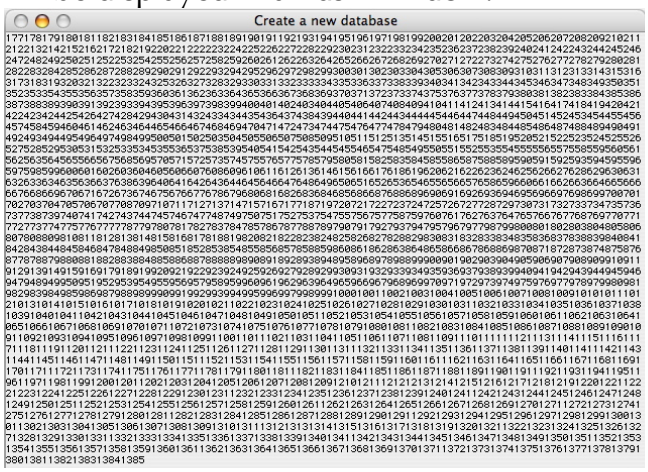
You now need to specify the exonic organization of your gene (how many exons and the first nucleotide position of each according to the reference cDNA).





Enter the gene symbol (please refer to the nomenclature system for correct gene symbol name http://www.gene.ucl.ac.uk/nomenclature/). You now need to import the reference cDNA sequence. For this select the corresponding text file created in the previous section. During the import process, all amino acids positions will be displayed in a flush window.

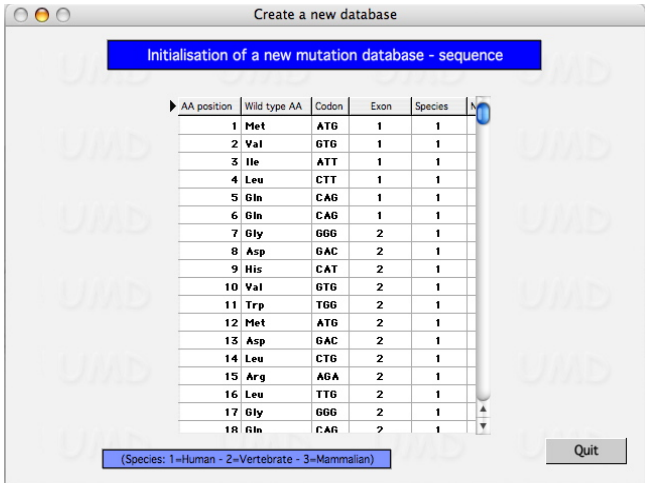




When the process is over, the coding reference sequence is displayed as a table. For each amino acid, the corresponding codon and exon will be displayed with the species value set to 1. The species value represents the conservation of the amino acid with the following code:

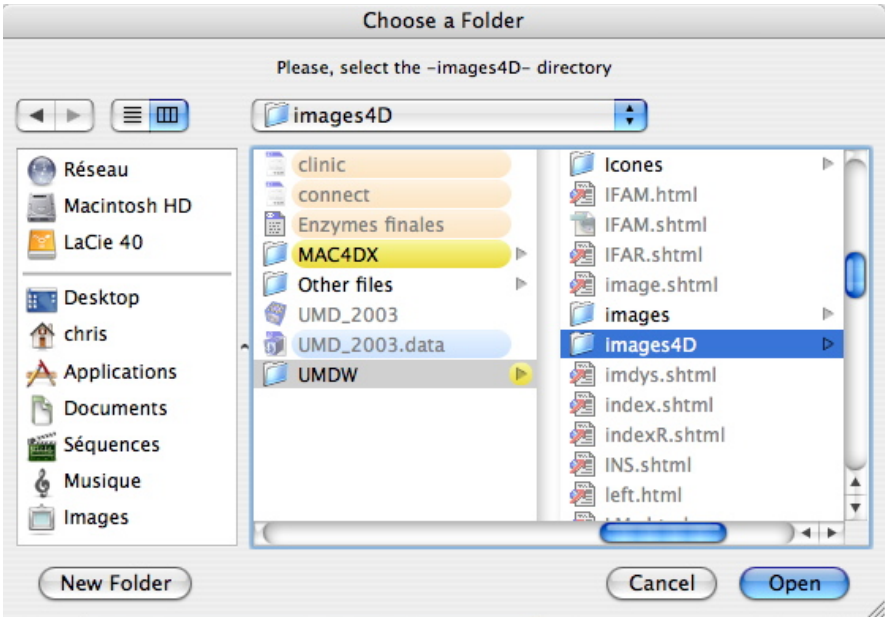
1 = AA only found in human
2 = AA conserved within mammals
3 = AA conserved within vertebrates

You can also set your own codes (4, 5 …). All data can be modified by a double click on the corresponding amino acid position.



Now click on the "Quit" button in order to validate all modifications. Re-launched the UMD software, all new data are now integrated and UMD® will generate the reference sequence of your gene and will create reference pictures such as the exon phasing and the reference sequence.

At this stage the UMD® software will ask you to specify the "images4D directory" position in your hard disk. It is located in the UMDW directory.

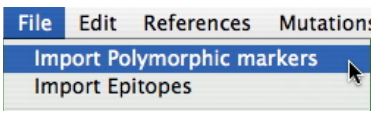


As UMD® is a generic software, it can be adapted to genes involved in cancer or genes involved in genetic diseases. To set-up this parameters please use the "***Gene type***" option from the "***File***" menu
The following message will appear "*This gene is involved in cancer. Do you confirm it?"* If your gene of interest is involved in cancer, just click "*Yes*" if not just click "*No*". If you incorrectly set this value to the wrong option, just select again the "***Gene type***" option from the "***File***" menu. The following message will appear: "*This gene is not involved in cancer. Do you confirm it?*" Just click "*Yes*" or "*No*".

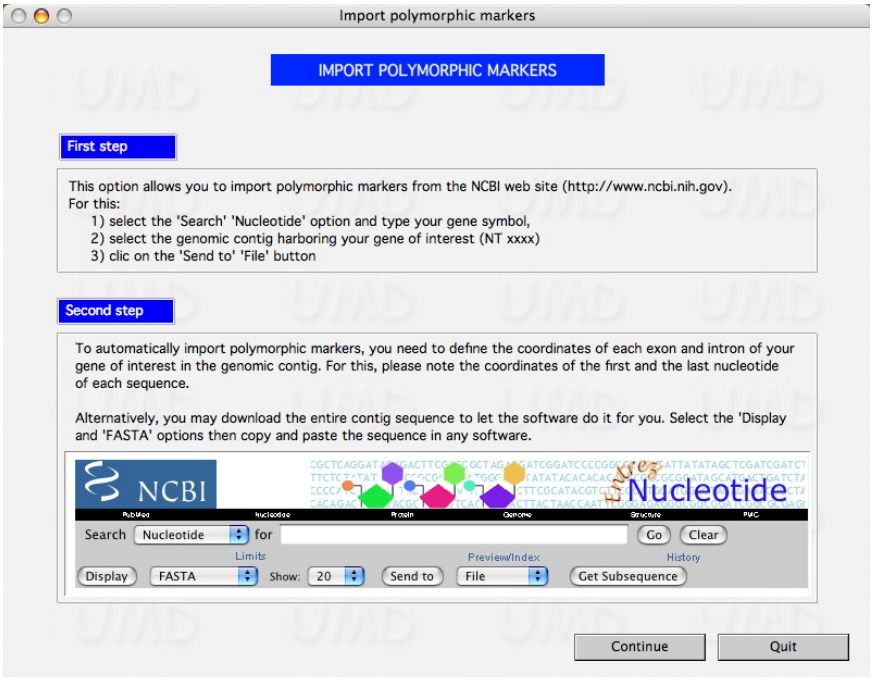
Your database is now ready to be used. If intronic sequences of your gene are known and localized in a genomic contig, we suggest that directly include intronic sequences. Nevertheless this process is optional and can be performed later.

II-b Integrate intronic sequences and SNPs from a genomic contig

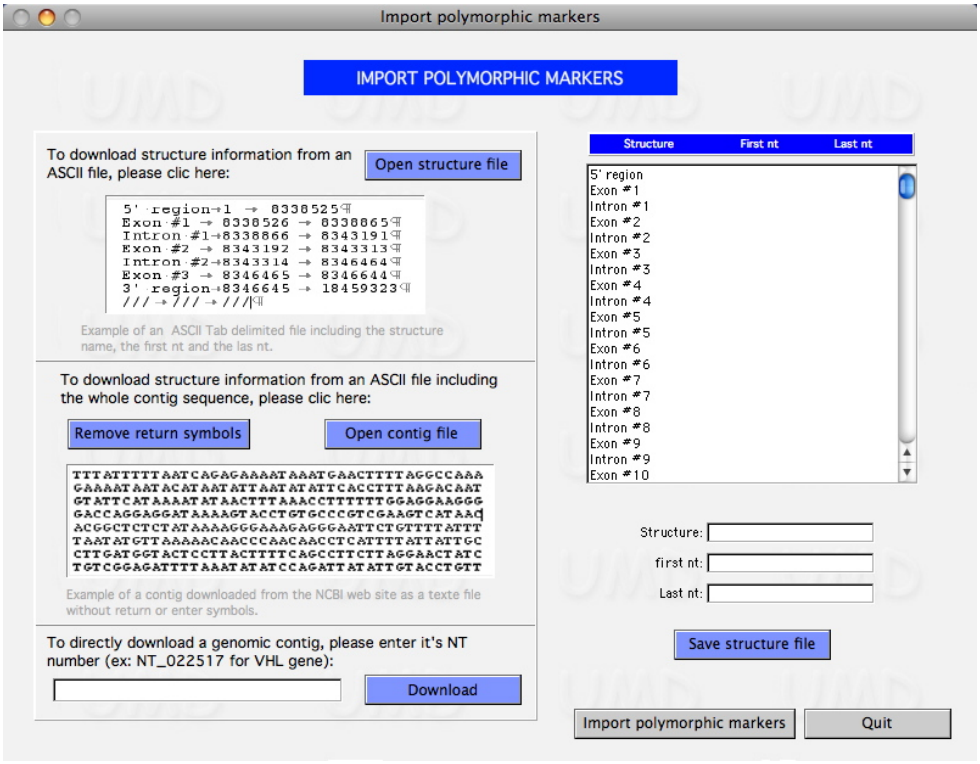
Select the "Import polymorphic markers" option from the file menu.



The following screen appears:



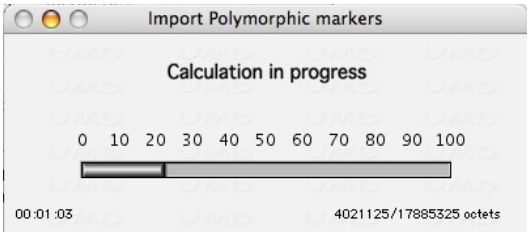
After downloading data from the NCBI (cf. II) click on “Continue”. A second screen appears that allows you to use three alternative ways to import the data:



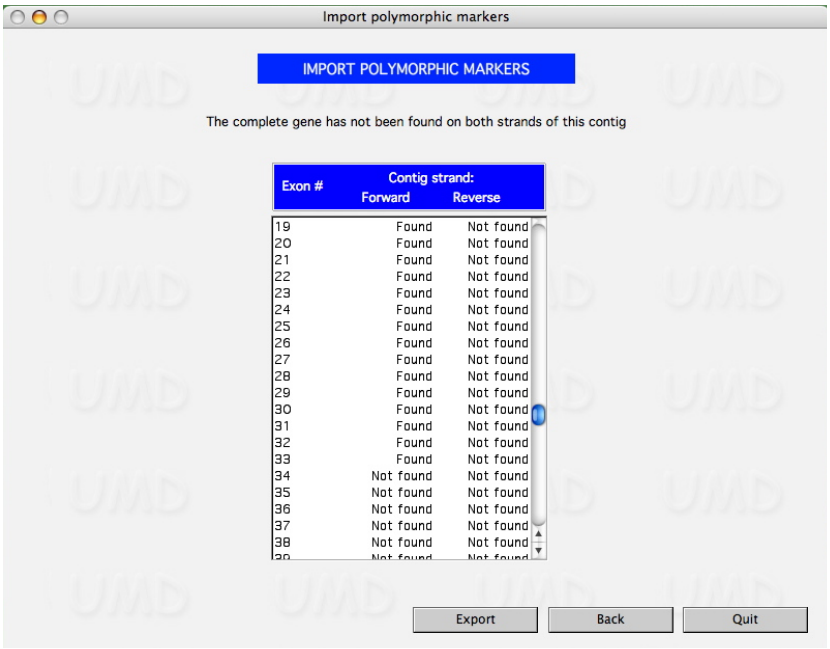
The first one is to use a tab-delimited file (ASCII file) describing the position of exons and introns in the contig (top). The second allows the import of the contig sequence file (only coding sequence) and the third one allows you to directly download through the Internet the contig using its NT code.

For option #2, if you got troubles to remove return symbols from your text editor software, you can use a specific tool to do that (press the remove return symbols button). If your text

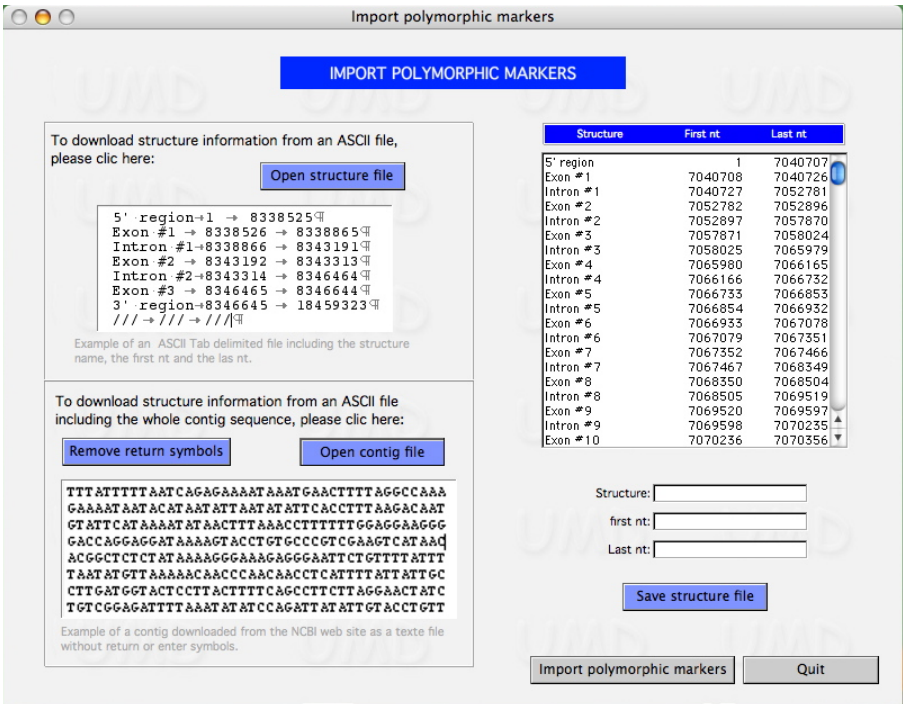
files are ready to use, press the “Open contig file” button and select the contig sequence text file. The software will then search on both strands of the contig for exonic sequences.



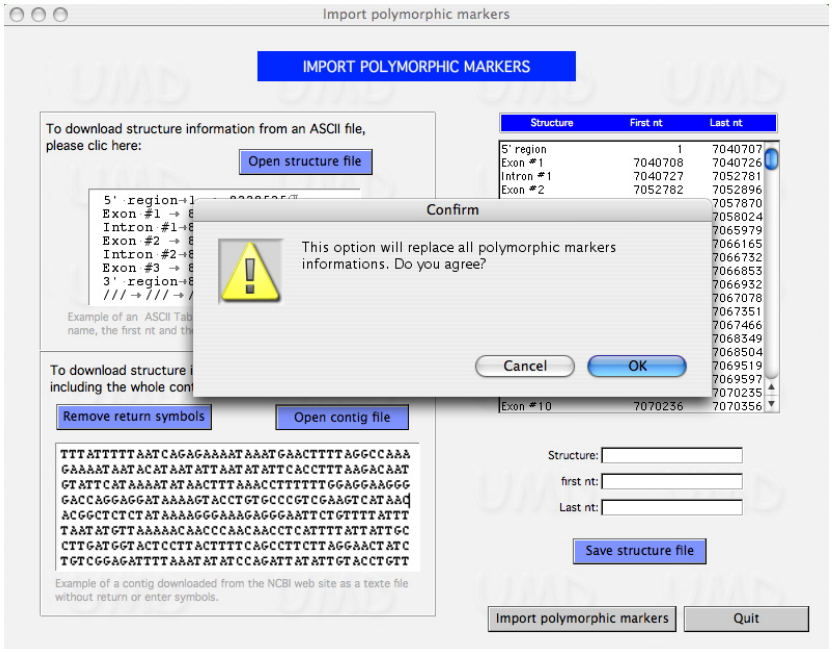
When this process is over UMD® can either have found all exonic sequences or only few of them. This is mainly due to polymorphisms in the contig sequence.



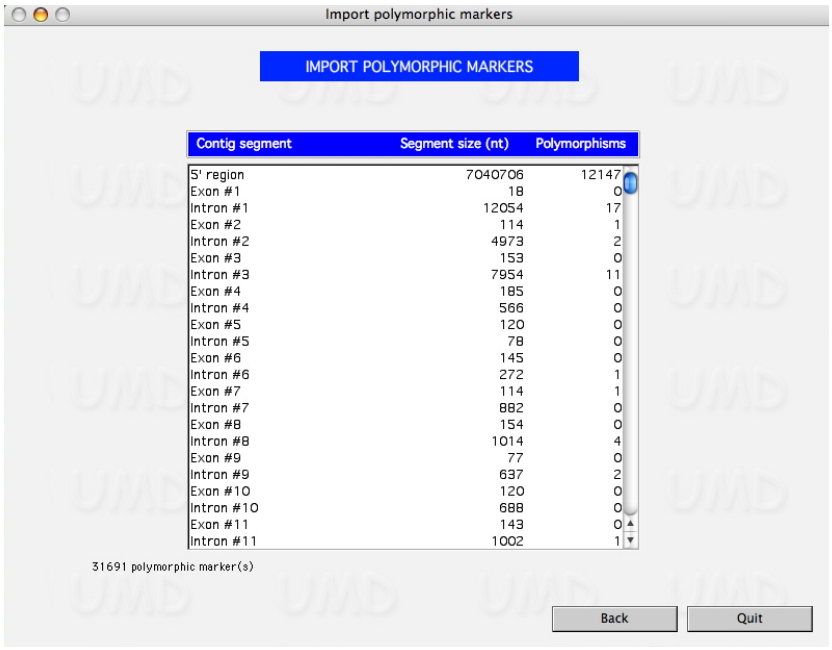
When no exon has been found, this screen displays partial results. You can export exonic sequences with the “Export” button and look for discrepancies between your coding reference sequence and the contig.
In this other example the UMD® software has found all exons on the forward strand of the contig. The intronic sequences are then automatically generated and included in UMD®.



You can now import polymorphic markers information with the corresponding button. A confirmation message will remind you that this process will update all information about polymorphic markers:



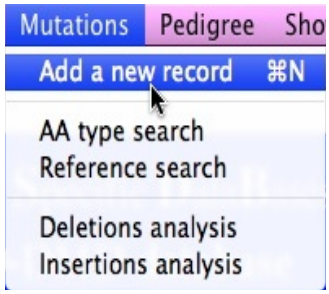
When the process is completed, a screen displays how many markers have been imported and their distribution in the various portions of the gene.



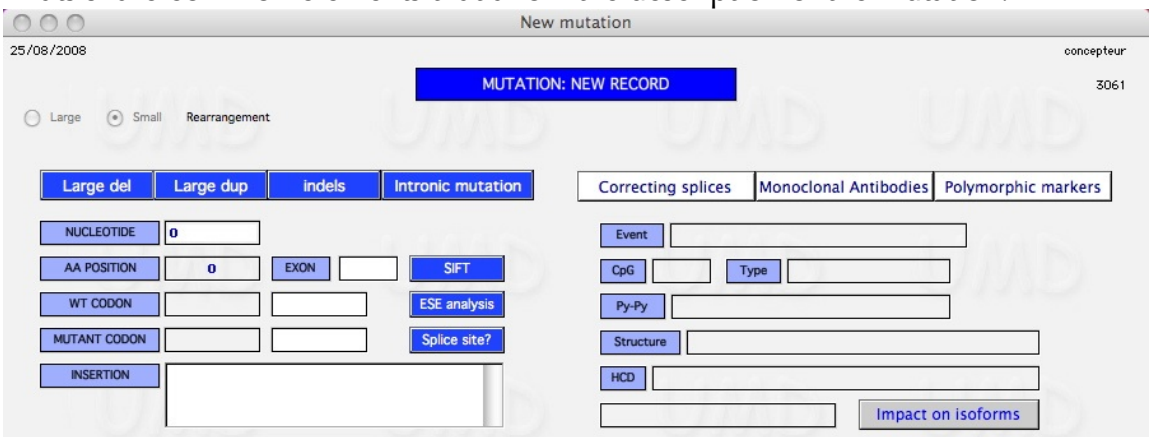
Note that even if you import manually non-coding sequences, you should use the naming rule: intron #xx and 5′ region or 3′ region. If you use other names, the software will not be able to import polymorphic markers neither to search for intronic reference sequences.

III- HOW TO ENTER A NEW MUTATION?

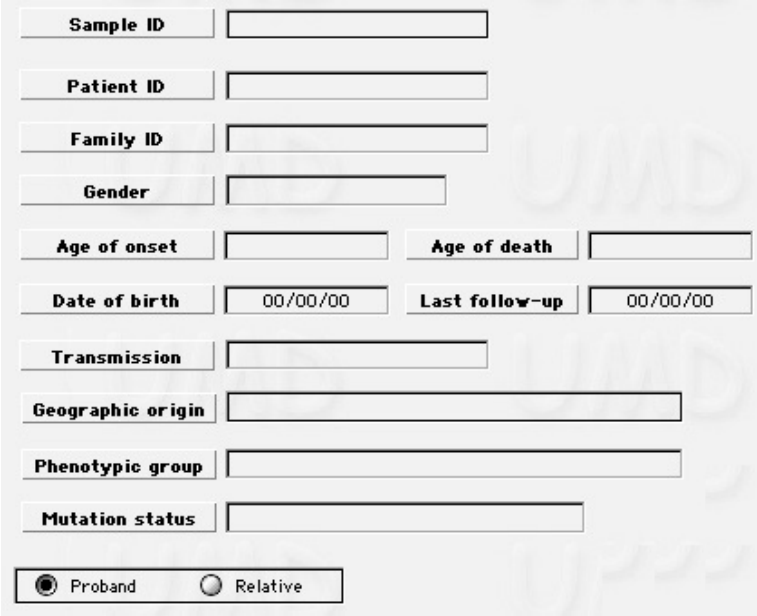
The UMD-database now includes the genetic code as well as the reference sequence of your gene of interest. Select in the "**Mutations**" menu the "**Add a new record**" option:



The UMD-database can have different input formats (contact us to get a specific design) but all formats share common elements that allow the description of the mutation:



The left part (A) allows the input of various types of mutation while the right part (B) includes data that are automatically computed by the software. Below this common part of the input screen, most formats also include the following fields that allow the input of information related to the sample and/or the patient:



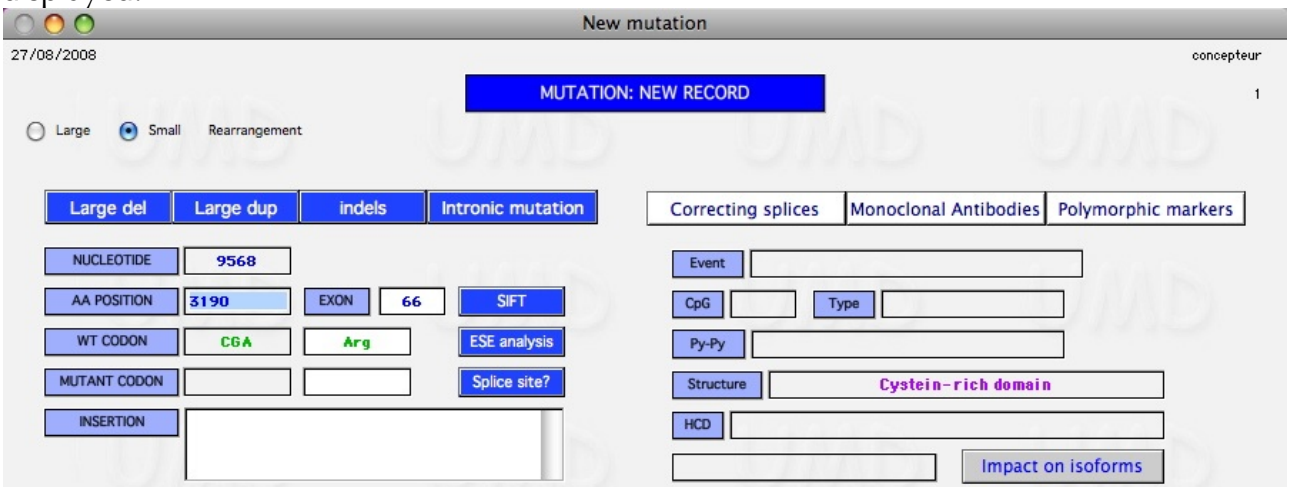
The "Sample ID" field is a unique identifier. You can set it using your own criteria. We suggest that you use a country code followed by a lab code a family and a patient code as well as a gender code. This will allow you to rapidly give enough information to the laboratory that reported the mutation to let it go back to the patient if necessary.

III-a - Point mutations

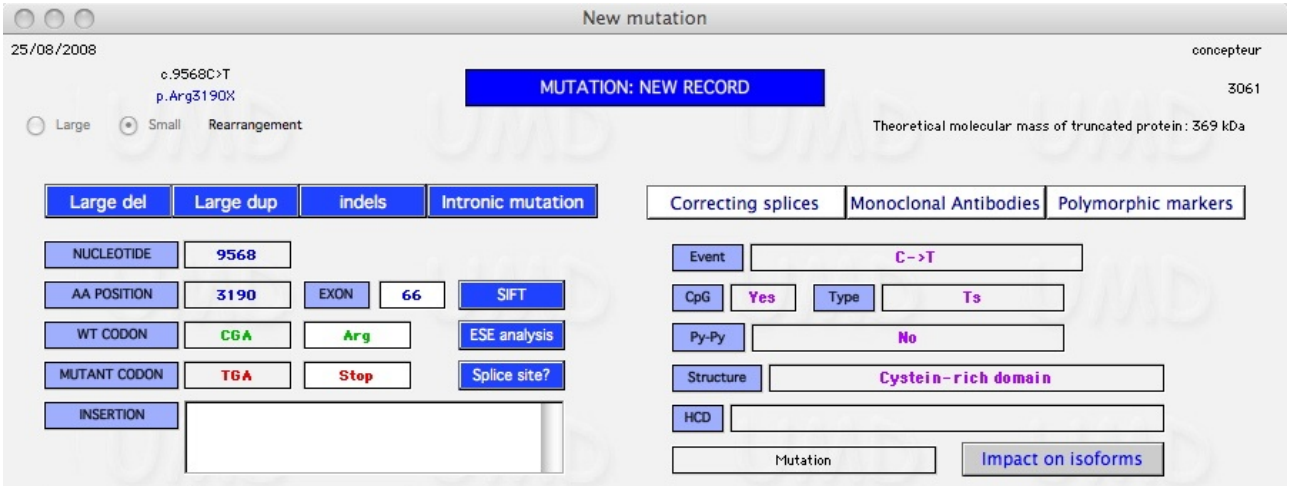
In the following examples, we will use the DMD gene (OMIM 300377) involved in Duchenne and Becker Muscular Dystrophies (cf. UMD-DMD LSDB).

1) c.9568C>T (p.Arg3190X)

In this example, we will show you how to enter a point mutation at nucleotide position #9568 resulting in an Arg3190X mutation. You can enter the mutation's position either at the nucleotide level (9568 in nucleotide field) or at the amino acid level (3190 in the AA position field).
The software will then display the wild type codon sequence (CGA), the corresponding wild type amino acid (Arg) and the exon where is localized this amino acid (Exon #66). Note also that the structural domain in which is localized the amino acid (Cystein-rich domain) is displayed.

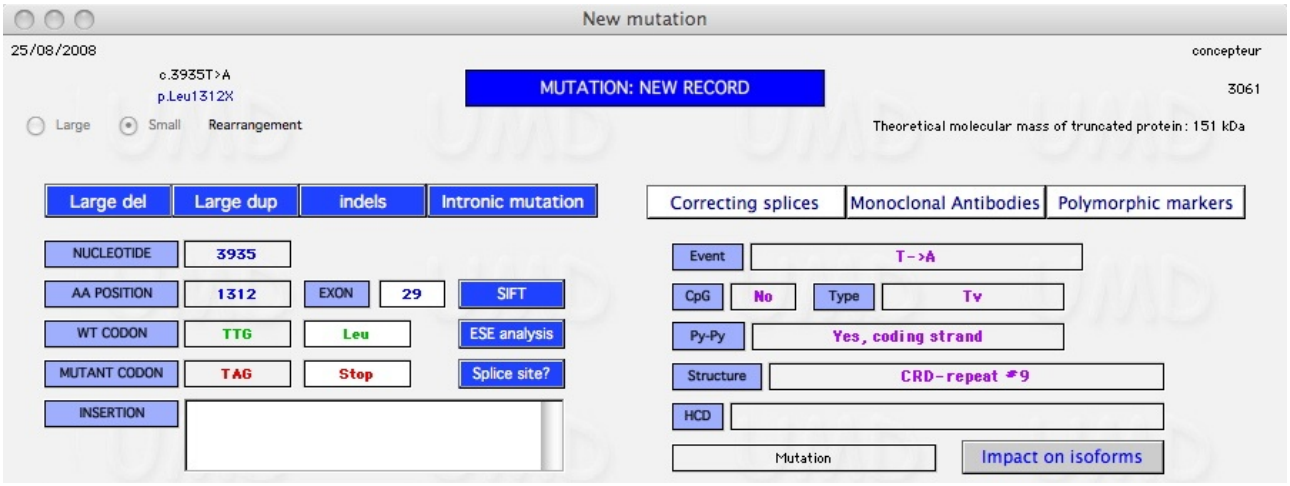


Now you need to enter the mutant codon "TGA". The software will automatically calculate and display the nomenclature name (c.9568C>T and p.Arg3190X) as well as the consequences of the mutation: "Molecular event = C->T"; "CpG = Yes"; "Mutation type = Transition (Ts)"; "Pyrimidine double (Py-Py) = No".

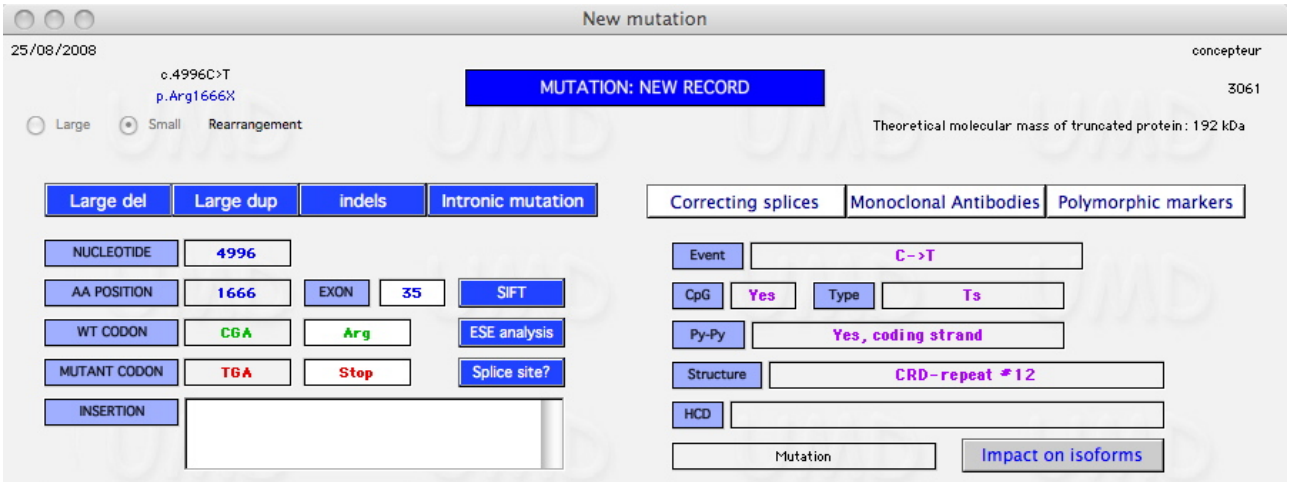


If you proceed with other nonsense mutations you should get the following results:

2) c.3935T>A (p.leu1312X)

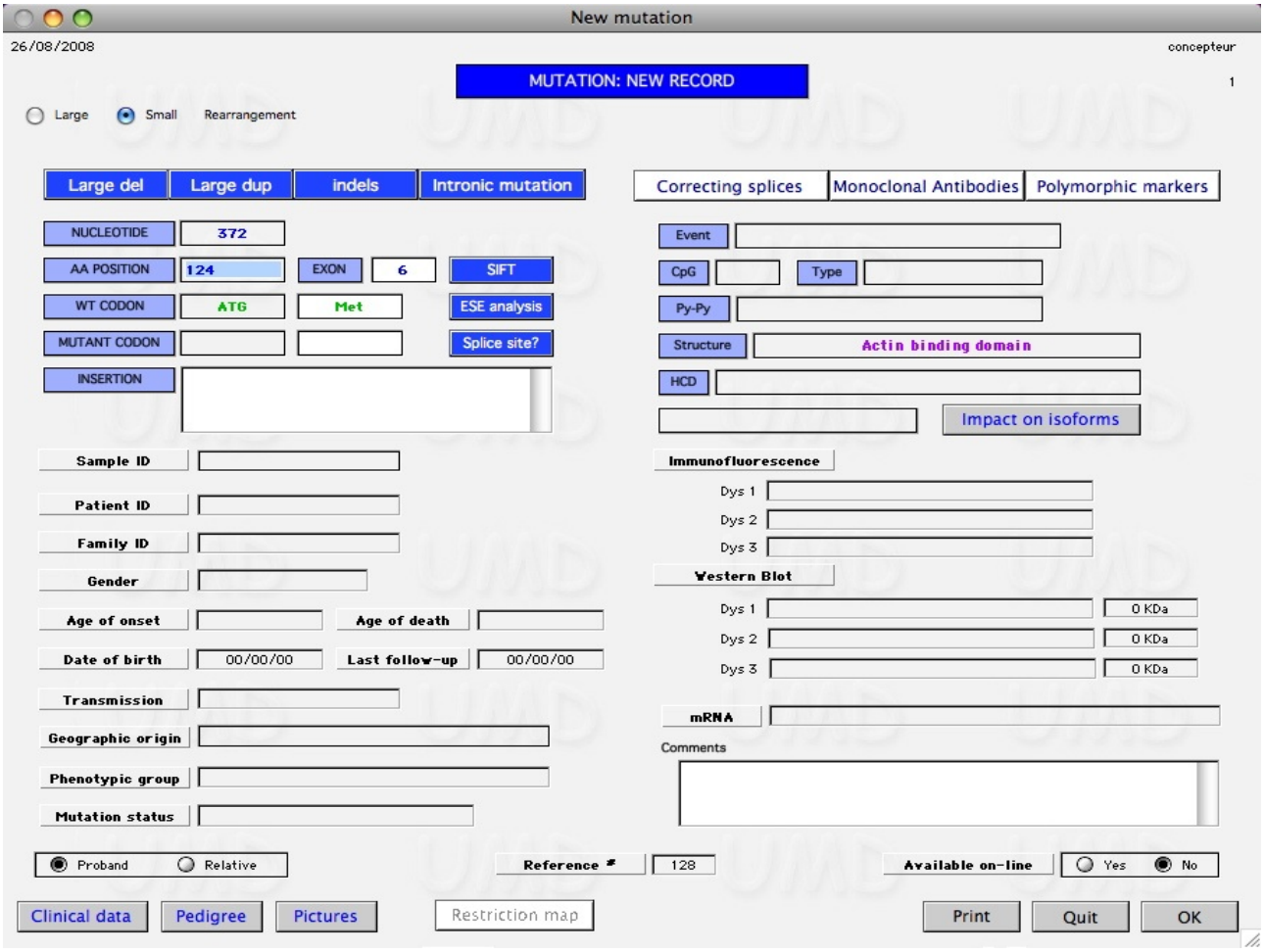


3) c.4996C>T (p.Arg1666X)



III-b- Small deletions

1) c.372delG



The mutation position selection is performed as for point mutations. To precisely define the deleted sequence you need to follow the UMD naming system. In this example you will create various deletions starting in codon #124 (first nucleotide is at the position 372).

Codon	A	T	G
Nucleotide position	372	373	374
UMD codon position	a	b	c

In the mutant codon section, you need to enter:

- The generic symbol “del”,
- The number of deleted nucleotides,
- The “UMD codon position” of the first deleted nucleotide.

If there is only one nucleotide deleted

- GTA ATG AAA -> GTA TGA AA del1a
- GTA ATG AAA -> GTA AGA AA del1b
- GTA ATG AAA -> GTA ATA AA del1c

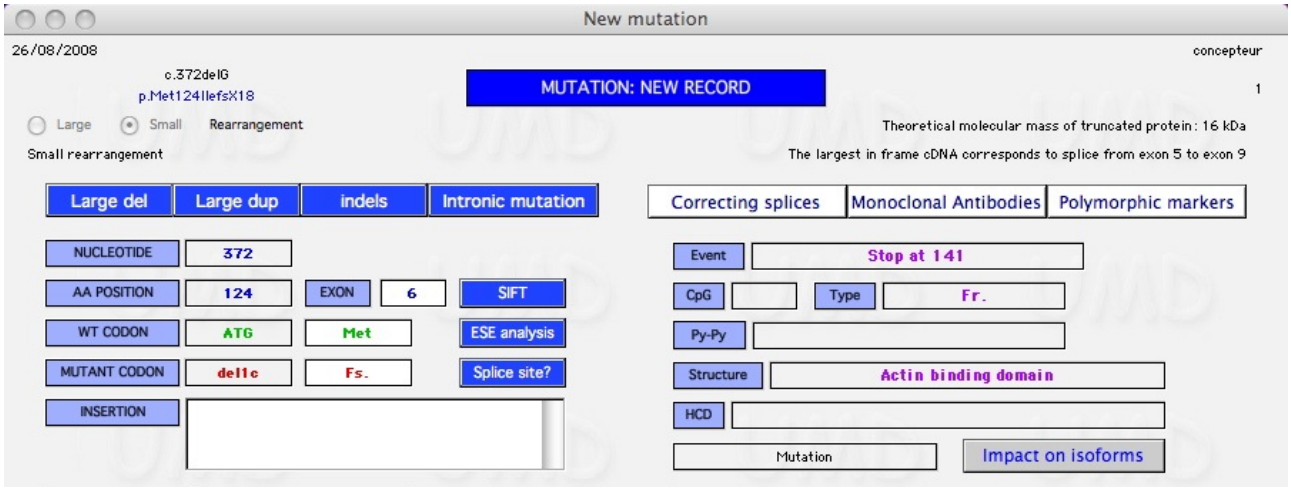
If there are two nucleotides deleted

- GTA ATG AAA -> GTA GAA A del2a

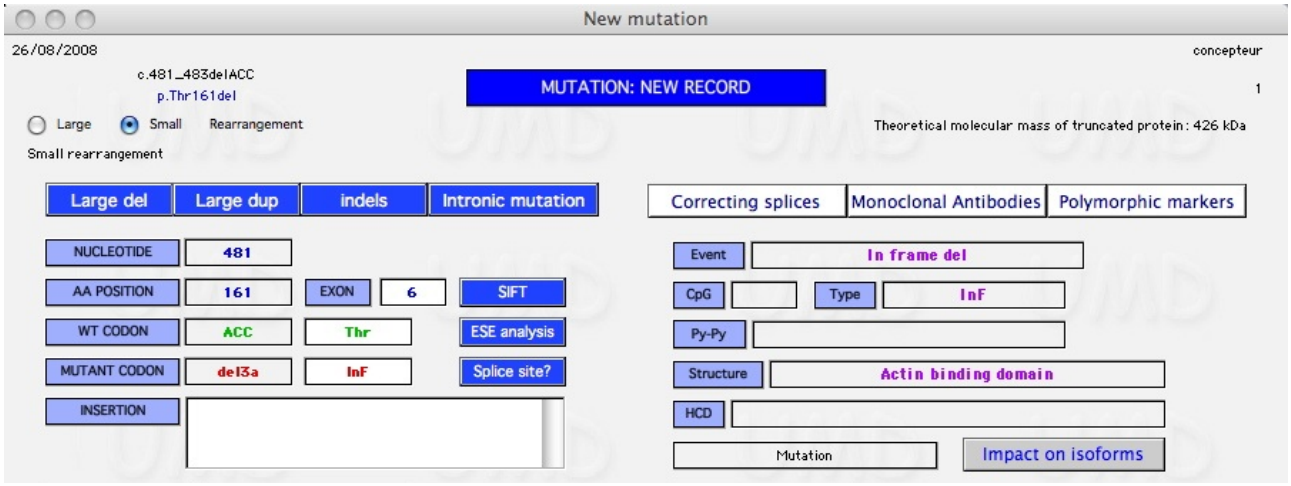
- GTA ATG AAA -> GTA AAA A del2b
- GTA ATG AAA -> GTA ATA A del2c
-

If there are three nucleotides deleted

- GTA ATG AAA -> GTA AAA del3a
- GTA ATG AAA -> GTA AAA del3b
- GTA ATG AAA -> GTA AAA del3c



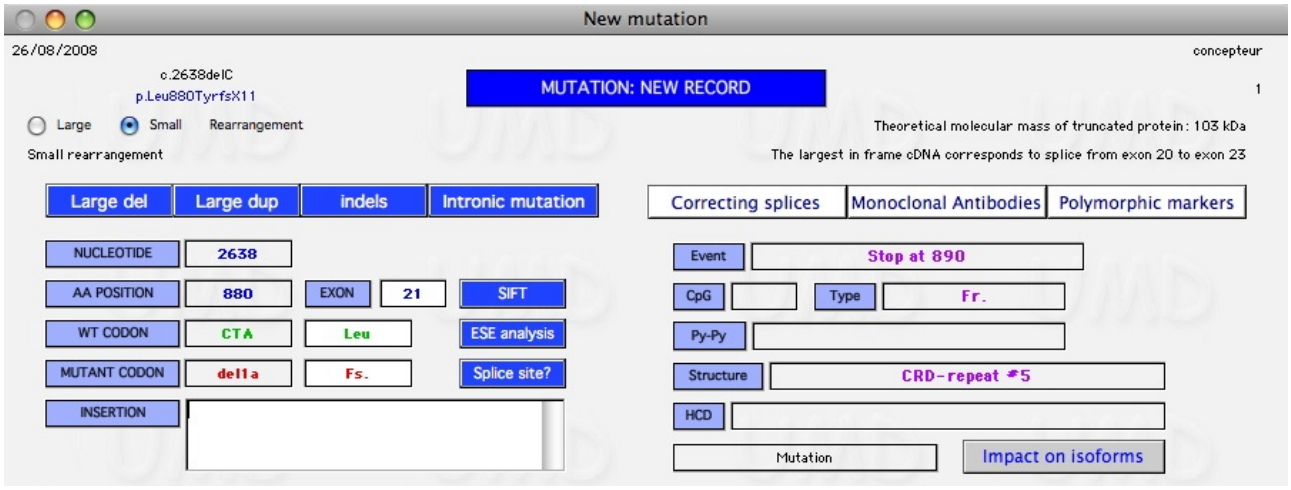
2) c.481_483delACC



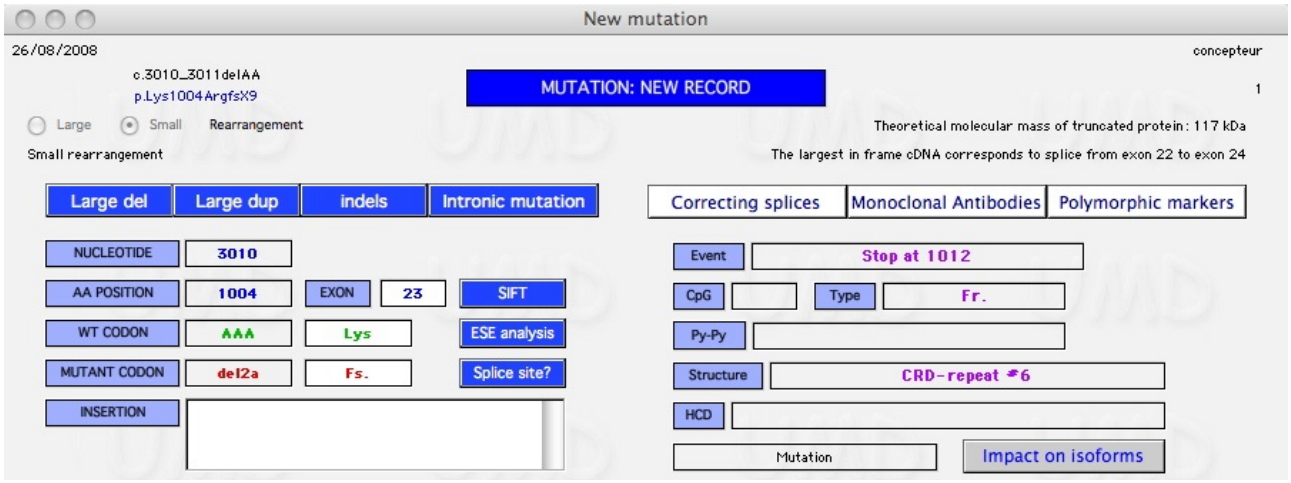
NB: Remember that the international nomenclature system states that for all descriptions the most 3' possible position is arbitrarily assigned to have been changed (especially when repeated sequences are involved).

Other examples are given below:

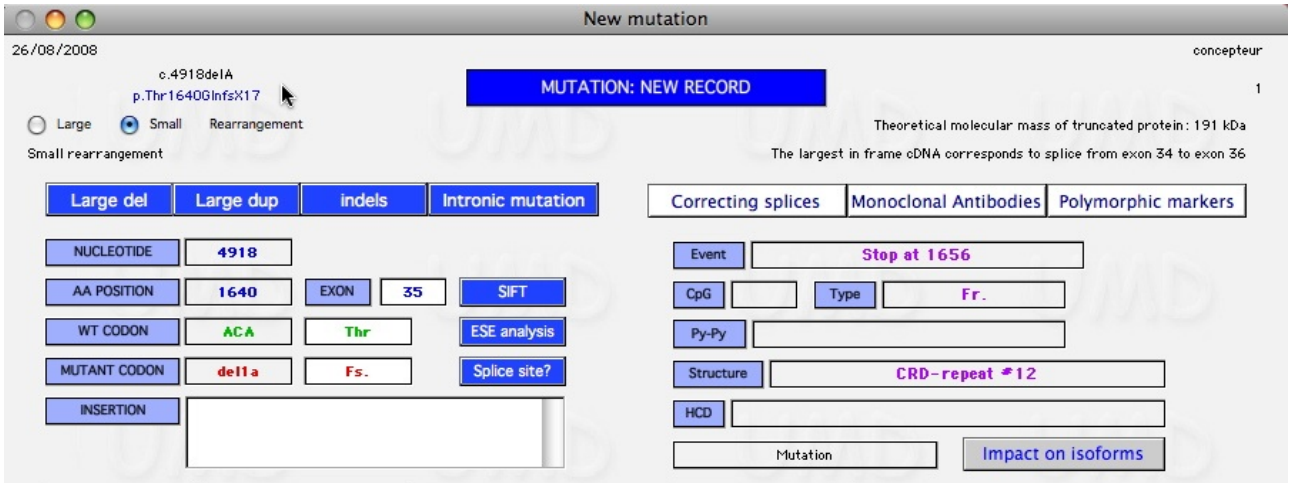
3) c.2638delC



4) c.3010_3011delAA

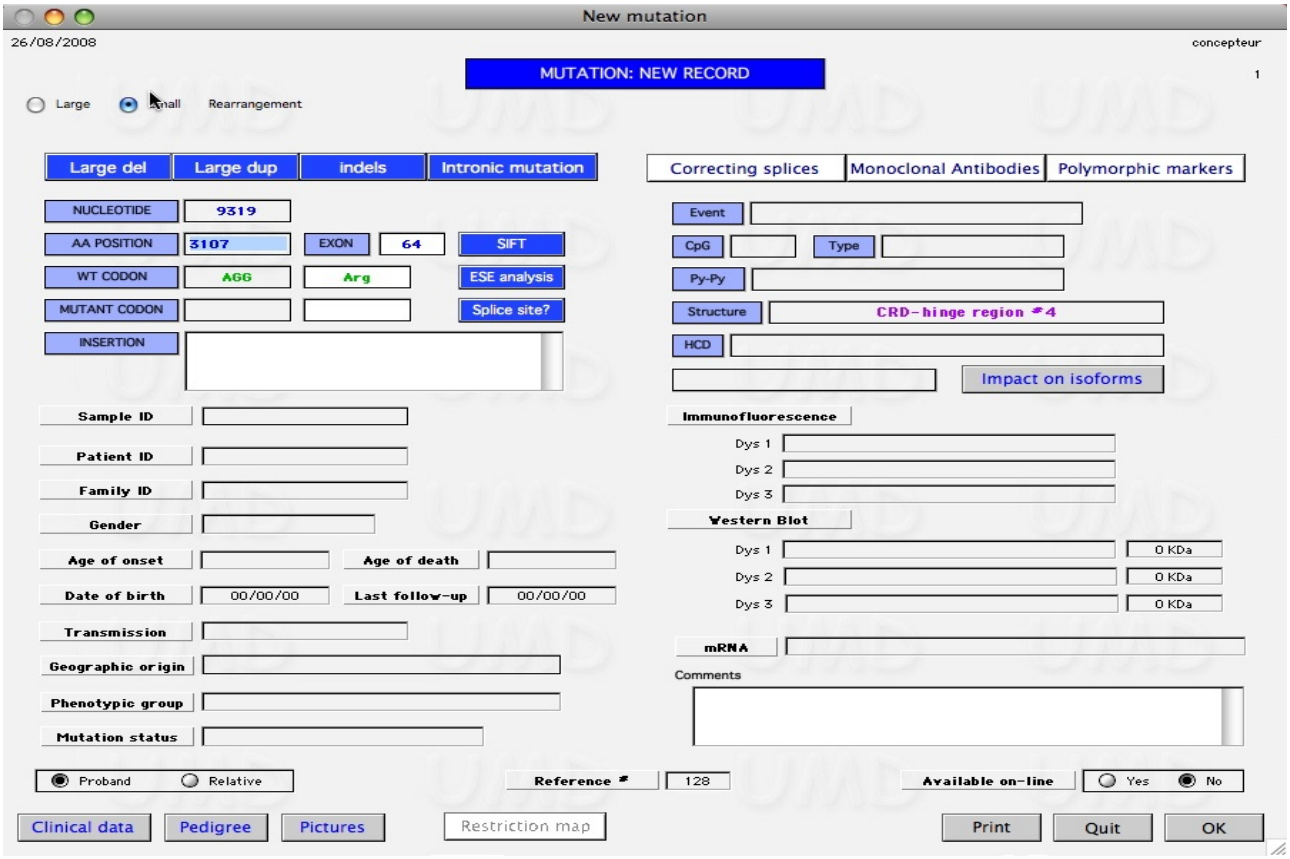


5) c.4918delA



III-c- Insertions

1) c.9319insC



The mutation position selection is performed as for point mutations. To define precisely where will be inserted the sequence you need to follow the UMD naming system:

Codon	A	G	G
Nucleotide position	9319	9320	9321
UMD codon position	a	b	c

In the mutant codon section, you need to enter:

- The generic symbol "ins",
- The number of inserted nucleotides,
- The "UMD codon position" of the **first nucleotide localized after the insertion**.

Insertion of one nucleotide

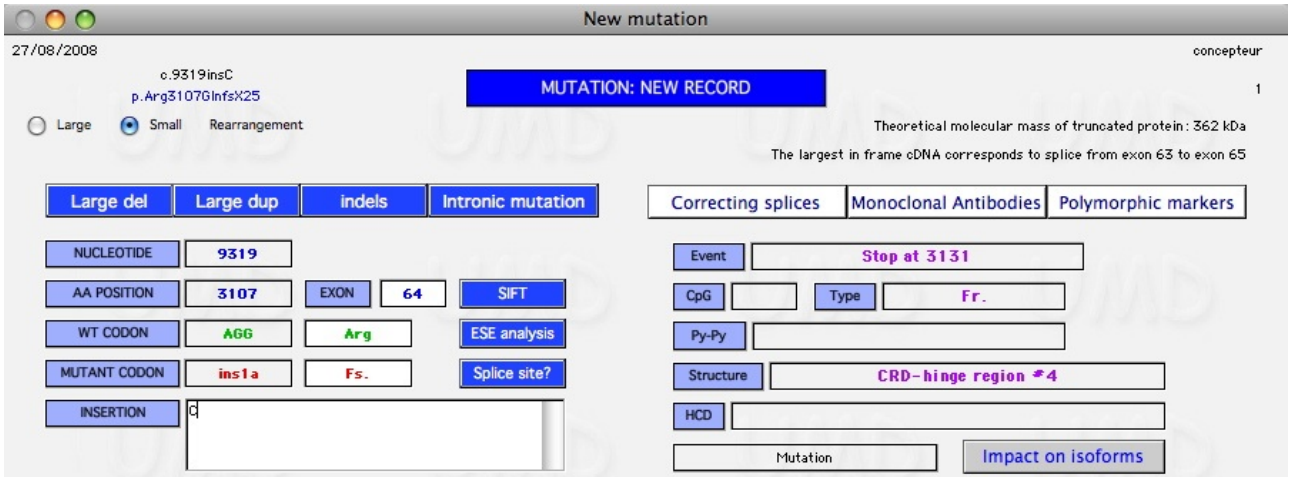
- AGG ACT GCC ins1a 1nucleotide (T) ATG GAC TGC C
- AGG ACT GCC ins1b 1nucleotide (T) AGT GAC TGC C
- AGG ACT GCC ins1c 1nucleotide (T) AGG TAC TGC C

Insertion of two nucleotides

- AGG ACT GCC ins2a 1nucleotide (TT) ATT GGA CTG CC
- AGG ACT GCC ins2b 1nucleotide (TT) AGT TGA CTG CC
- AGG ACT GCC ins2c 1nucleotide (TT) AGG TTA CTG CC

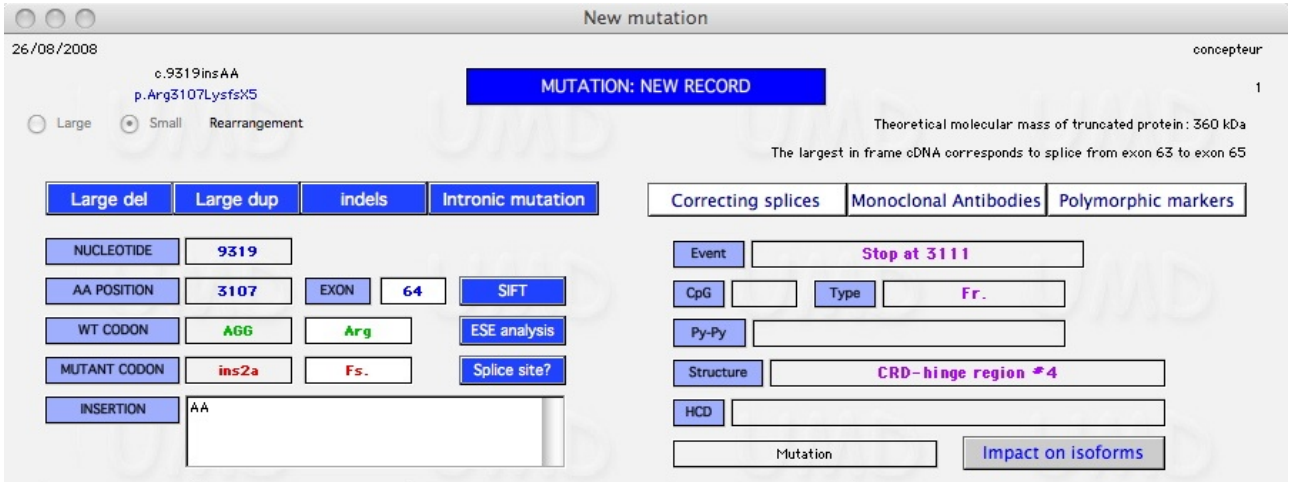
Insertion of three nucleotides

- AGG ACT GCC ins3a 1nucleotide (TTT) ATT TGG ACT GCC
- AGG ACT GCC ins3b 1nucleotide (TTT) AGT TTG ACT GCC
- AGG ACT GCC ins3c 1nucleotide (TTT) AGG TTT ACT GCC

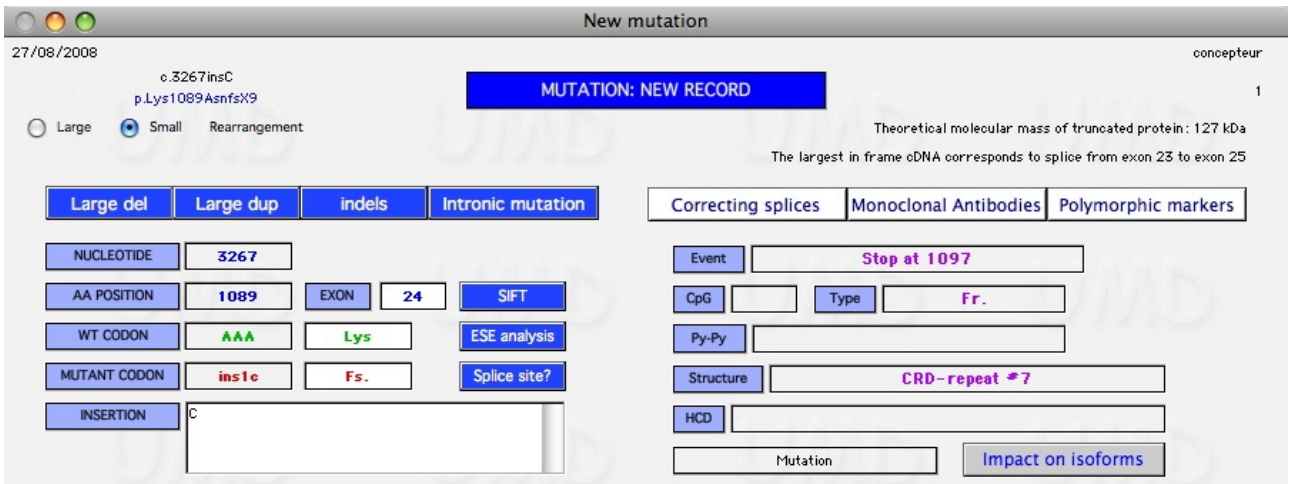


Other examples are given below:

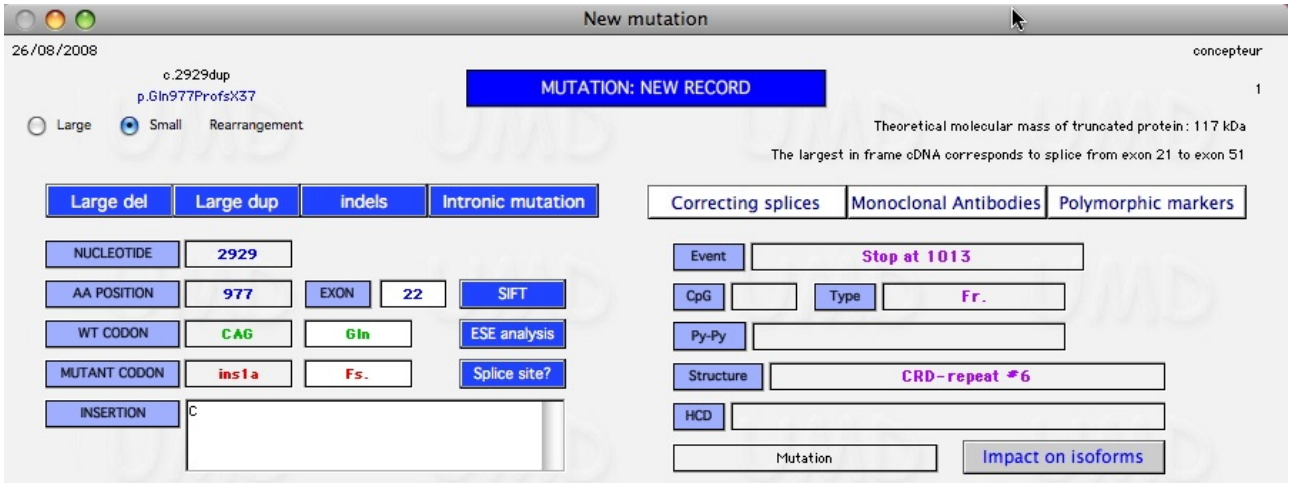
2) c.9319insAA



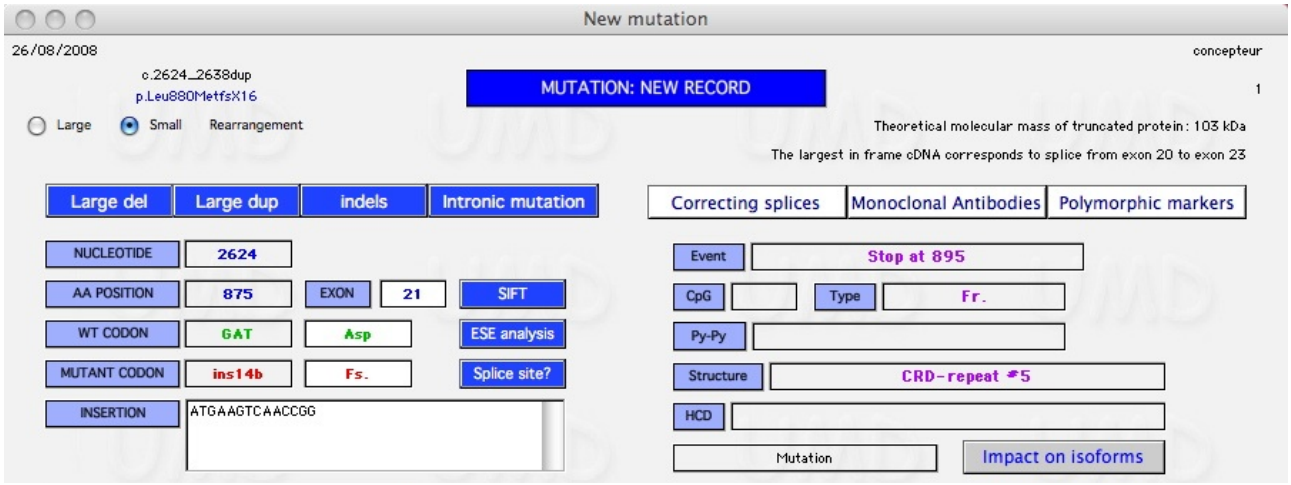
3) c.3267insC



4) c.2929dup

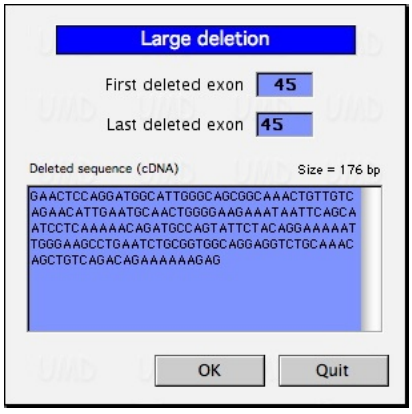


5) c.2624_2638dup

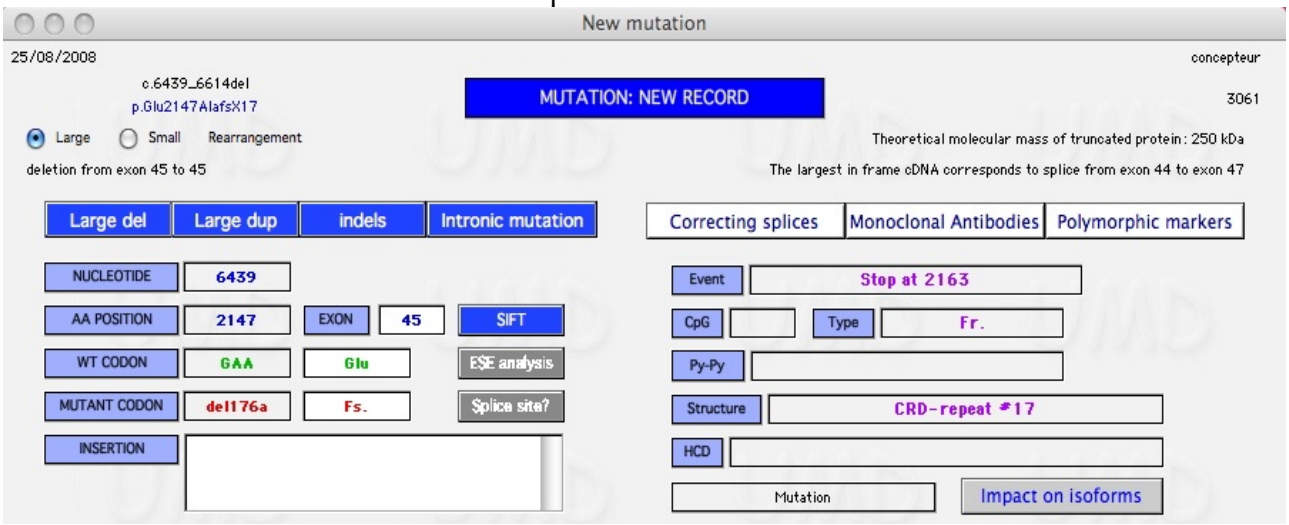


III-d- Large deletions

Large deletions correspond to the deletion of one or more exons of the gene. In order to facilitate the description of large deletions, you can use the "***Large del***" button. This will display a specific window where you can specify the first and last deleted exons. The software shows the deleted cDNA sequence:

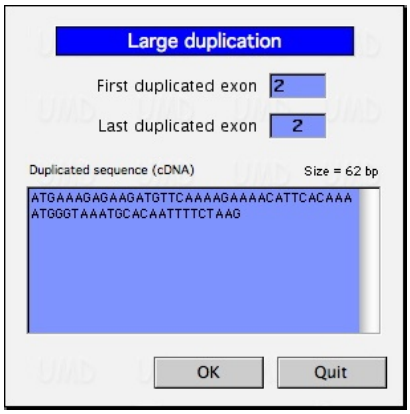


After validation, the software will fill all fields and calculate the position of the premature termination codon when a frameshift is produced:

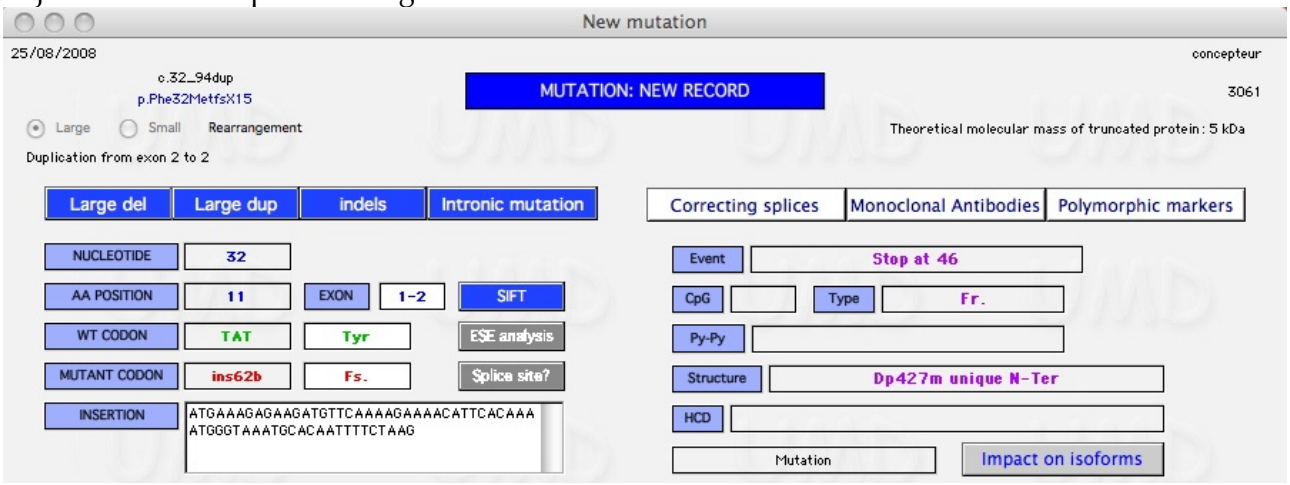


III-e- Large duplications

As for large deletions, large duplications correspond to the duplication of one or more exons of the gene. In order to facilitate the description of large duplications, you can use the "***Large dup***" button. This will display a specific window where you can specify the first and last duplicated exons. The software shows the duplicated cDNA sequence:



After validation, the software will fill all fields and calculate the position of the premature termination codon when a frameshift is produced. This implies that the duplication if adjacent to the duplicated segment and in the same orientation:

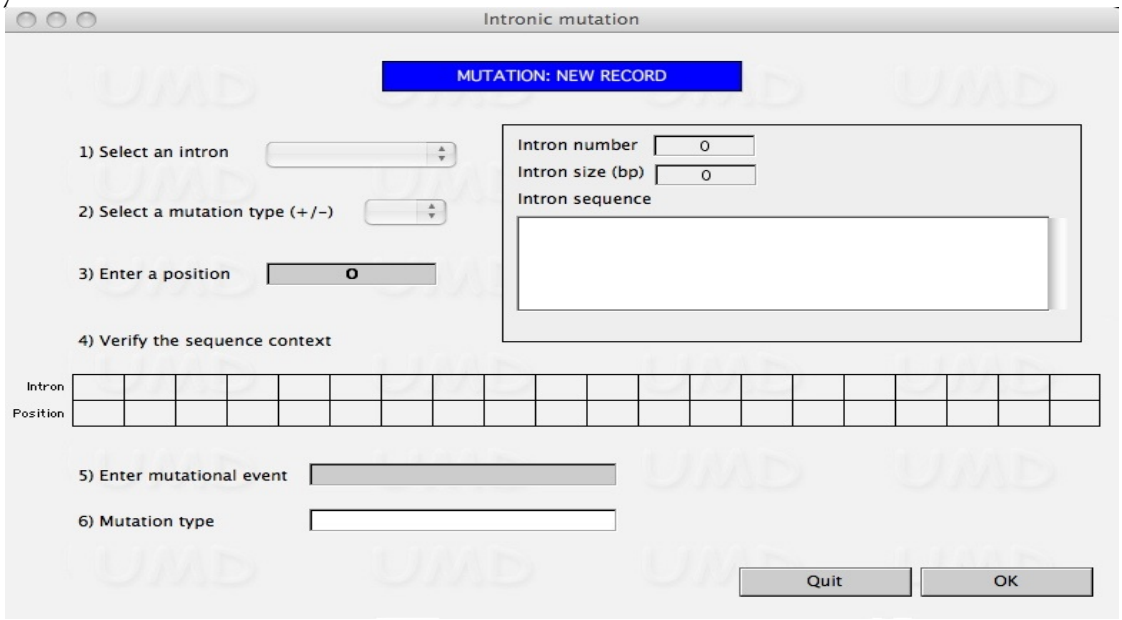


III-f- Intronic mutations

Beware that you can only enter intronic mutations (also called in UMD splice mutations) if you have previously defined the reference sequence of each intron using the proper nomenclature (cf. chapter 2). Because in most clinical diagnostic laboratories the naming system for such mutations is the alternative nomenclature (IVS), UMD integrate both nomenclatures (for more see HGVS website).

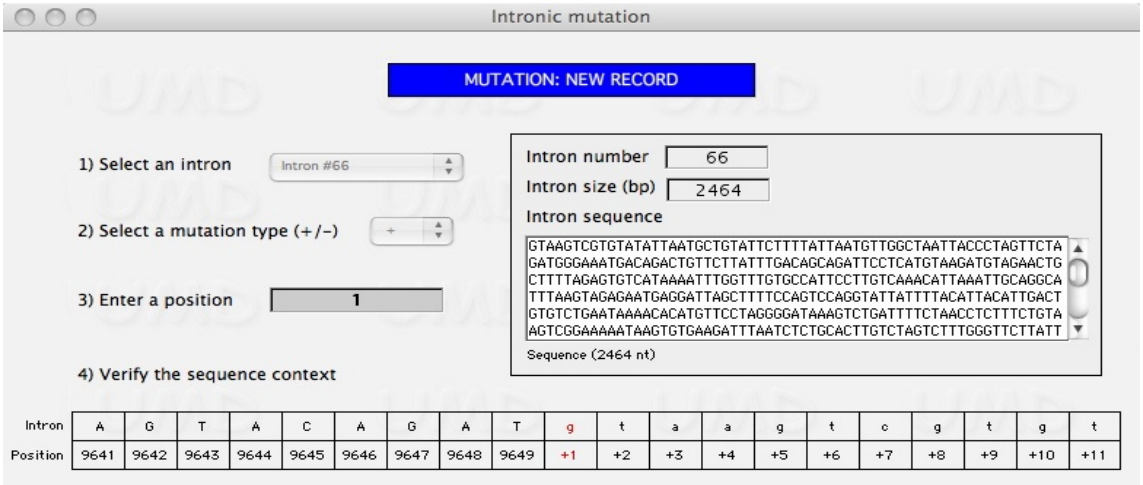
1) c.9649+1G>A (c.IVS66+1G>A)

You can enter an intronic mutation using the "***Intronic mutation***" button. A new window is displayed:



To enter a mutation you need to process in three steps. For example, for a mutation of the first nucleotide of the intron (position +1):

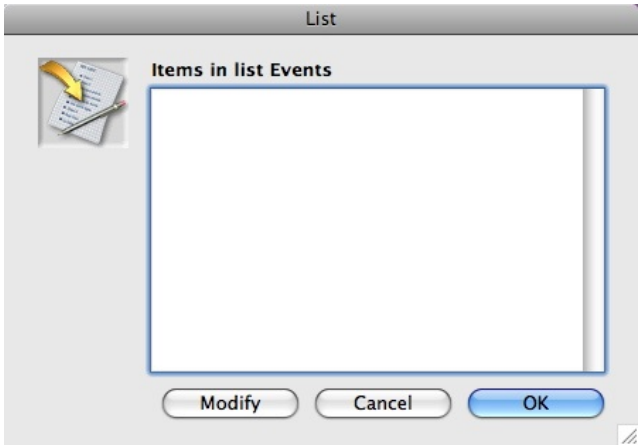
- Select the intron # (here #66)
- The mutation type (+ for mutation of a donor splice site, - for an acceptor splice site)
- The intronic position (here #1)



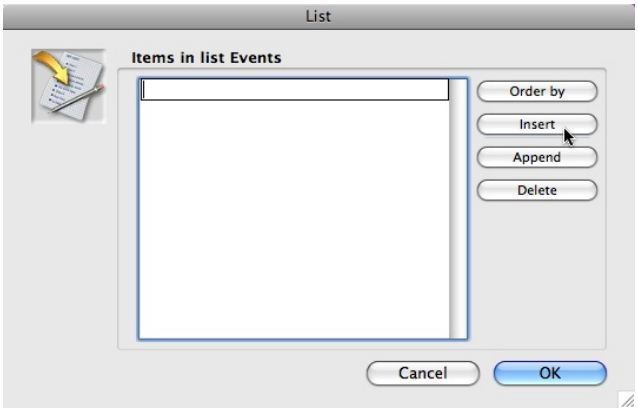
The UMD® software will display at the bottom of the window the reference sequence surrounding the selected position. Intronic sequence is displayed in small letters and exonic sequence in capital letters.

A window appears to allow a rapid selection of the mutational event:

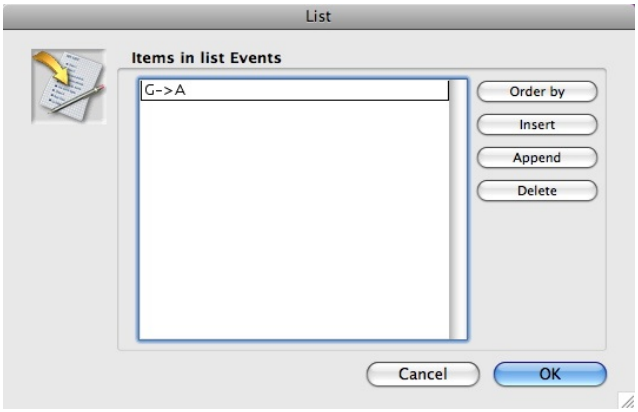
If the corresponding event is not available in the list, click the "**Modify**" button. At this stage, you can add, remove, insert or delete an event:



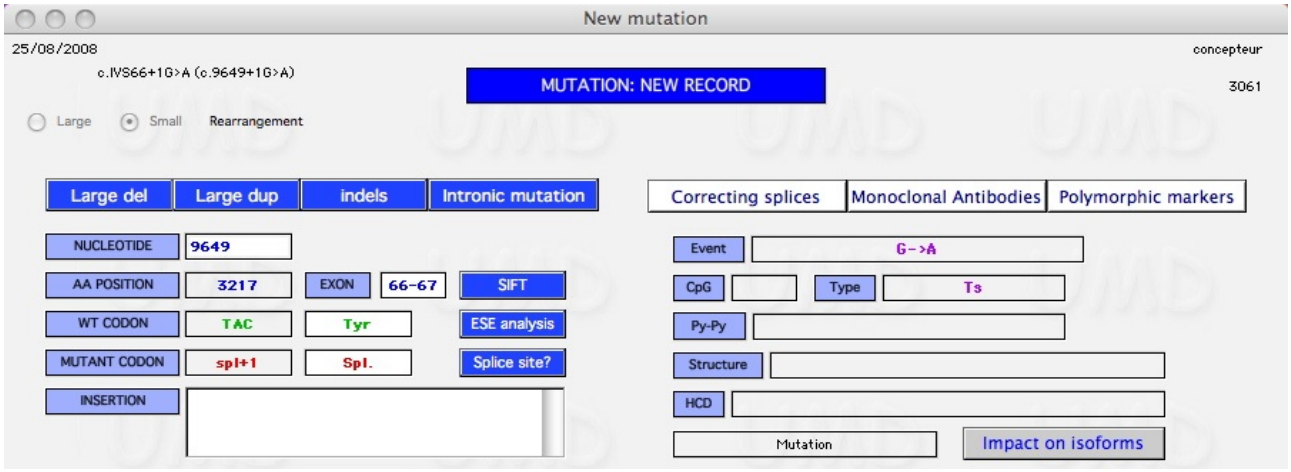
For the mutation c.9649+1G>A, the molecular event is a transition G to A. Click on the "***Insert***" button and add "**G->A »** in the text area :



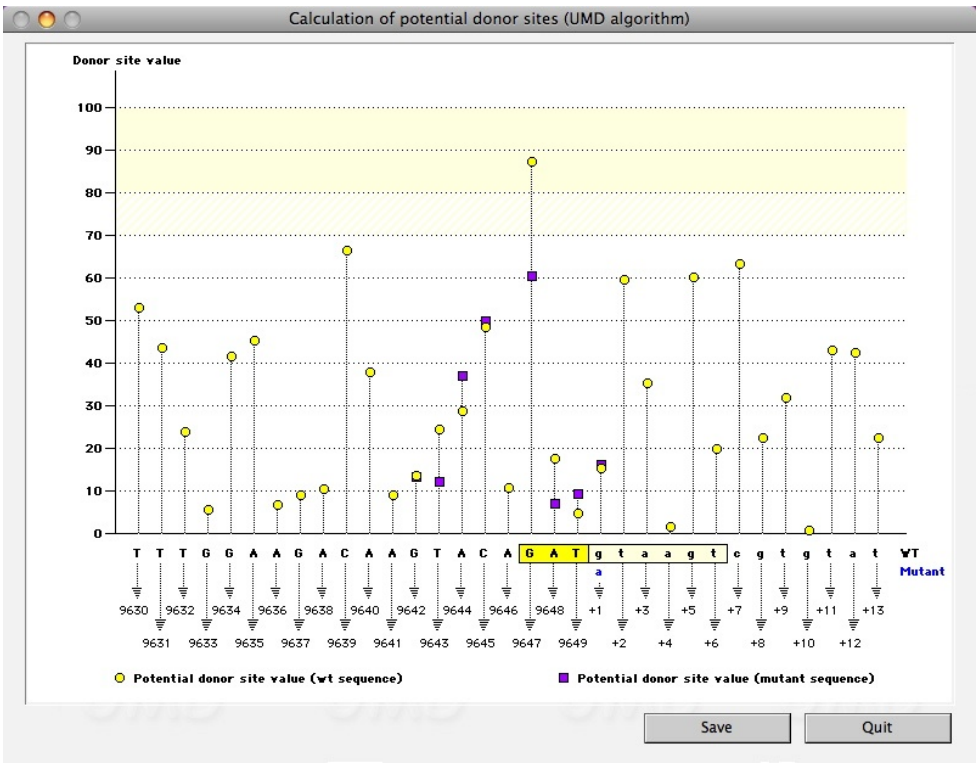
Then press the "***OK***" button



The international name of the mutation is automatically calculated by UMD®. Note that for intronic mutations it also includes the alternative naming system as mentioned before.



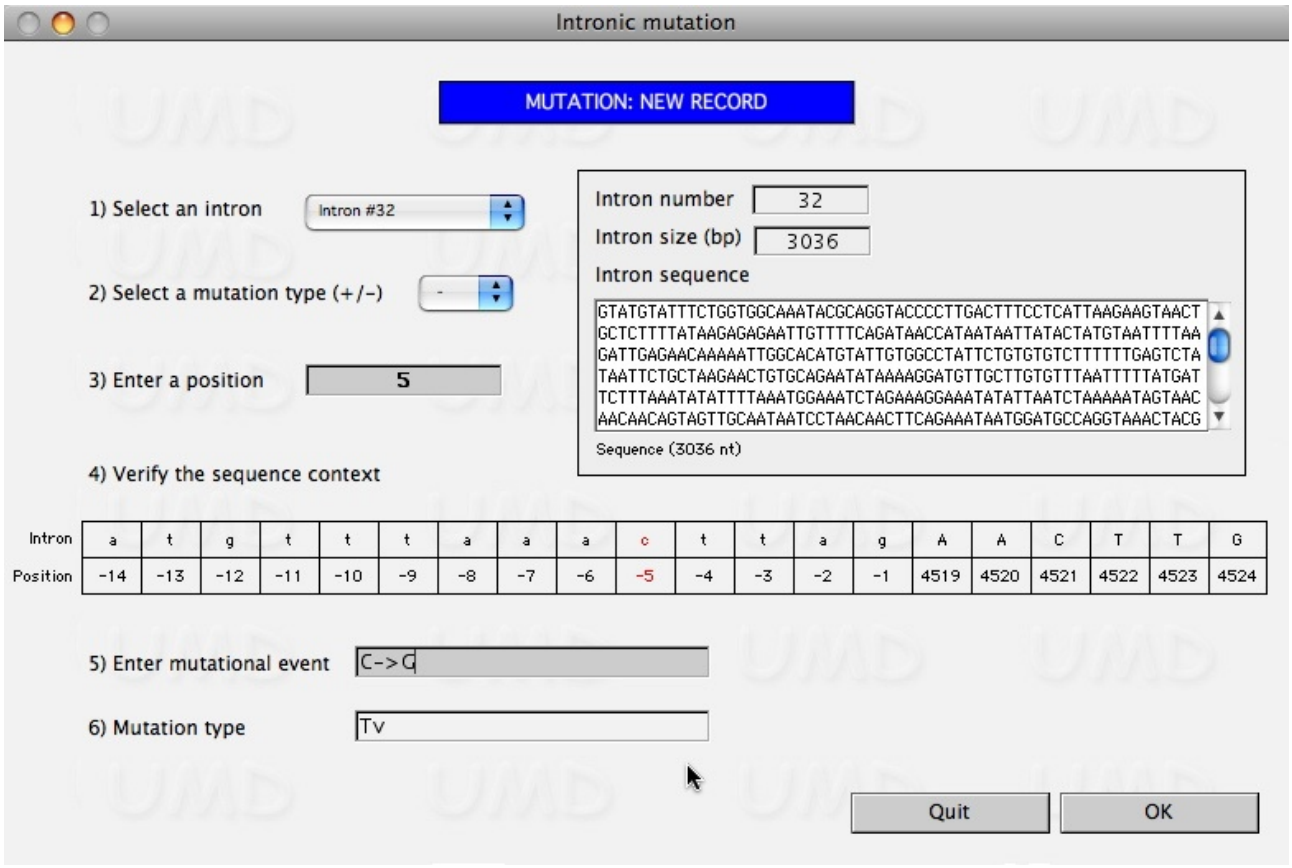
For all intronic mutations, two windows appear to display the possible impact of the mutation on donor (example below) and acceptor splice sites.
Exonic sequence are displayed in capital letters with corresponding nucleotide numbers (9630 to 9649) while intronic sequence are displayed in small letters with corresponding nucleotide numbers (+1 to +13). The normal donor splice site is shown in yellow (dark yellow for exonic nucleotides, light yellow for intronic nucleotides). The CV for each potential donor splice site (nine-nucleotide sequence) in the vicinity of the mutation is displayed according to the UMD rule (80-100 strong splice site; 70-80 = weaker splice site; 0-70 not a splice site).
Wild-type sequence CV are displayed as yellow square and mutant sequence CV as purple square. Each square represents the CV of the potential splice site beginning with the corresponding nucleotide.



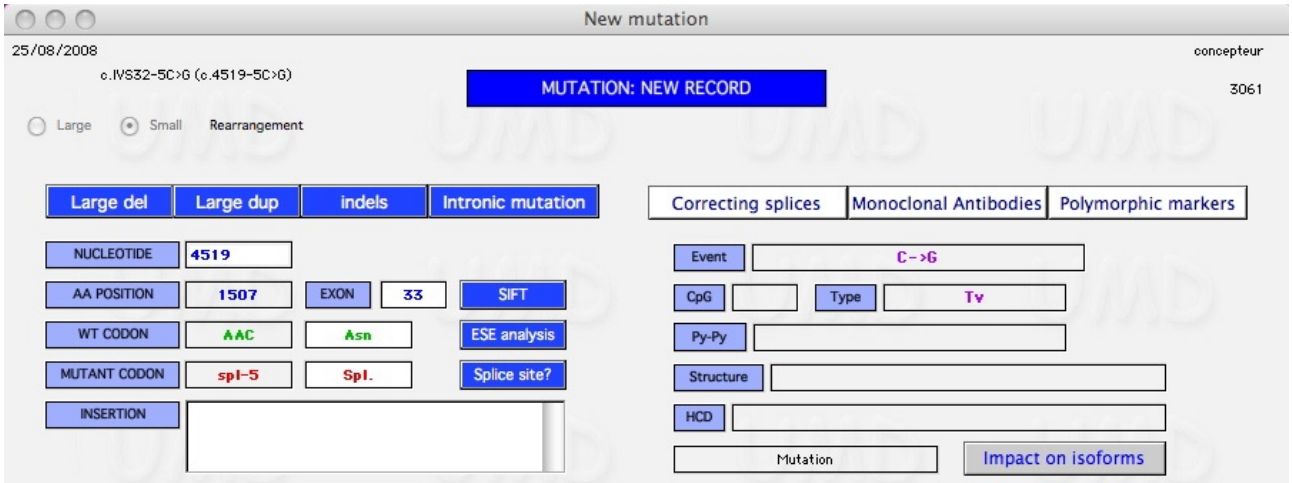
Disruption of a donor splice site

2) c.4519-5C>G (new nomenclature) / c.IVS32-5C>G (old nomenclature)

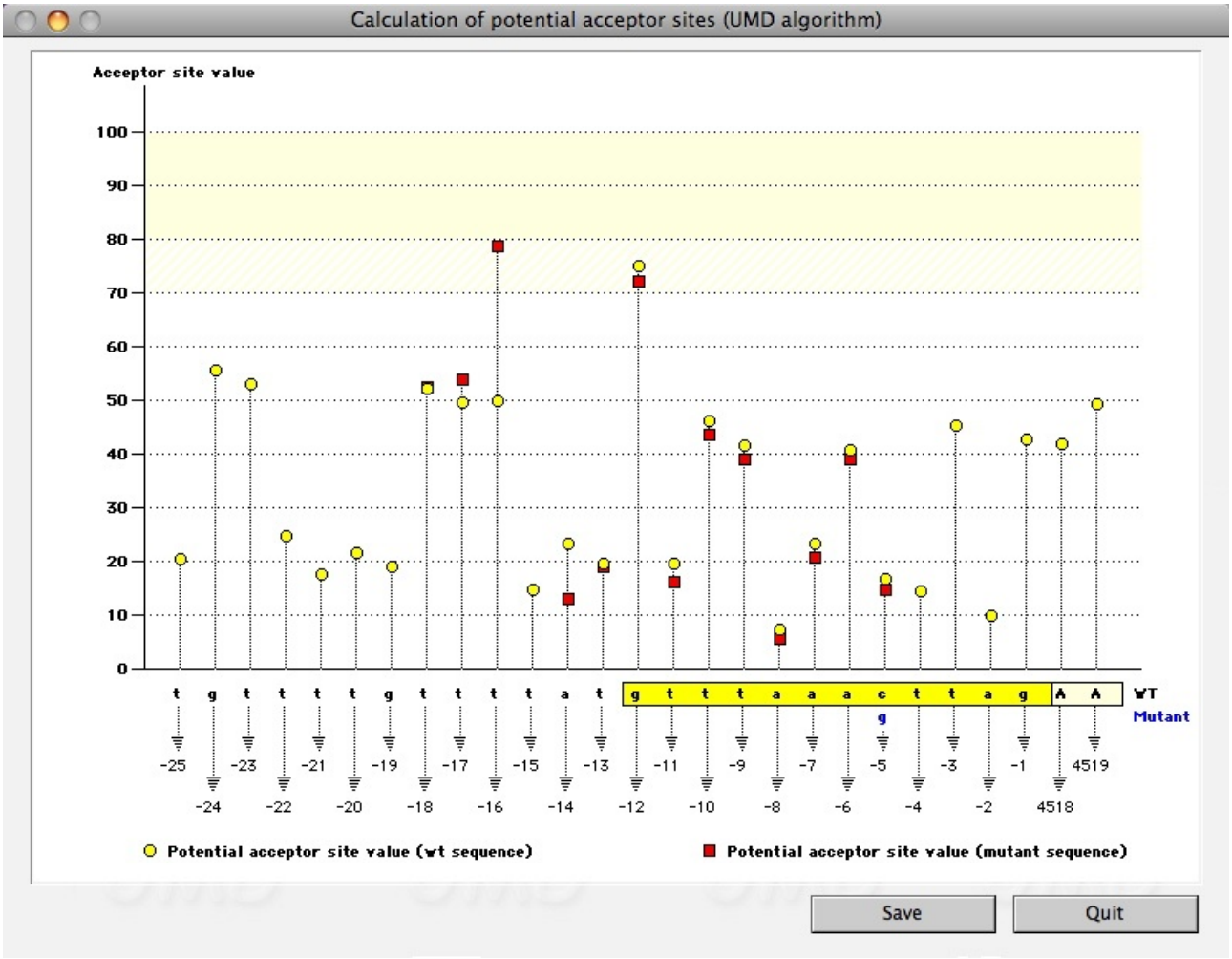
To enter this mutation, proceed as in the previous example, you should get:



After validation of these data, UMD® calculates the consequences of this mutation and give it a name:



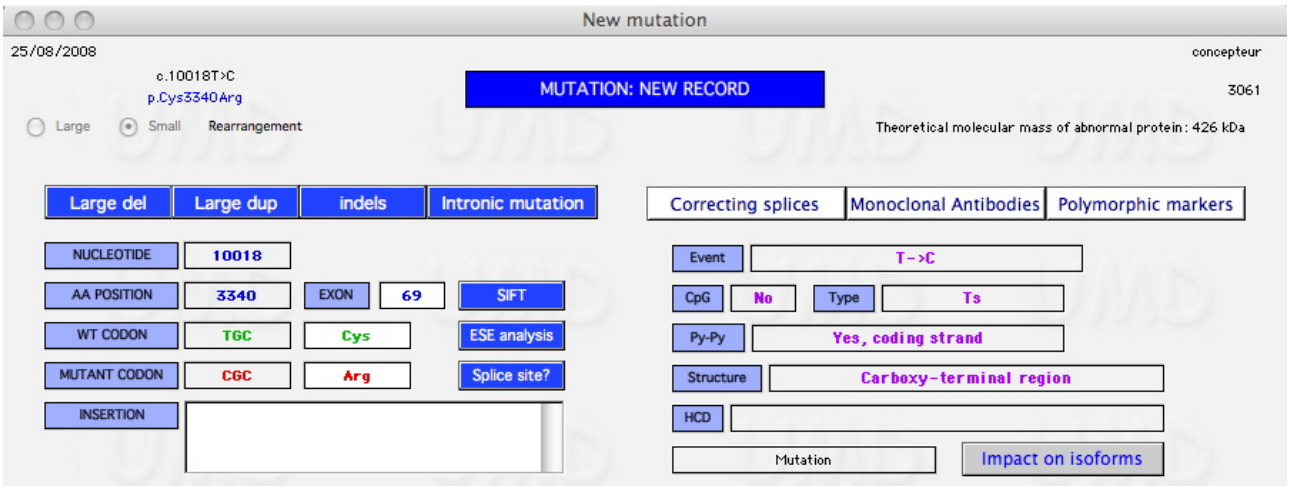
Here the mutation at position -5, does not affect significantly the wt acceptor site (value drops from 75 to 72, pink circle) but instead creates a cryptic splice site (CV = 78.95) beginning at position -16 (wt site begins at position -12) leading to the incorporation of 4 nt and therefore to a premature stop codon and presumably to NMD (experimentally demonstrated).


Calculation of potential acceptor sites (UMD algorithm)
Acceptor site value
100
90
80
70
60
50
40
30
20
10
0
t g t t t t g t t t t a t g t t t a a a c t t a g A A
WT
Mutant
g
-25 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 4518 4519
Potential acceptor site value (wt sequence)
Potential acceptor site value (mutant sequence)
Save
Quit

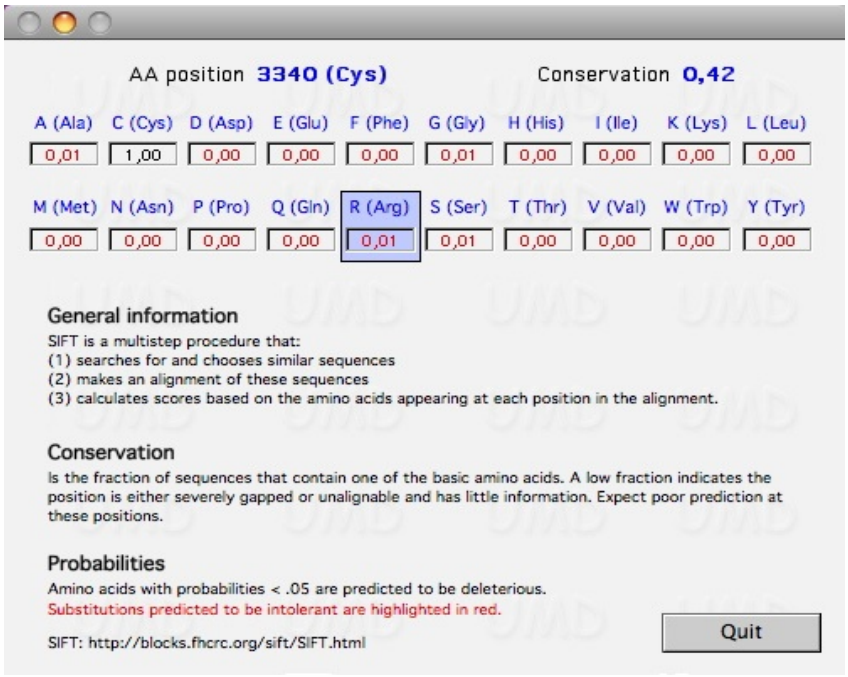
III-g- Missense mutations

To enter missense mutations, select the position using either the nucleotide or the codon number and enter the mutant codon. The software will automatically calculate the mutational event, the involvement of a CpG and/or Pyrimidine doublet, the structural domain that harbors the mutation as well as the exon and the mutant amino acid.

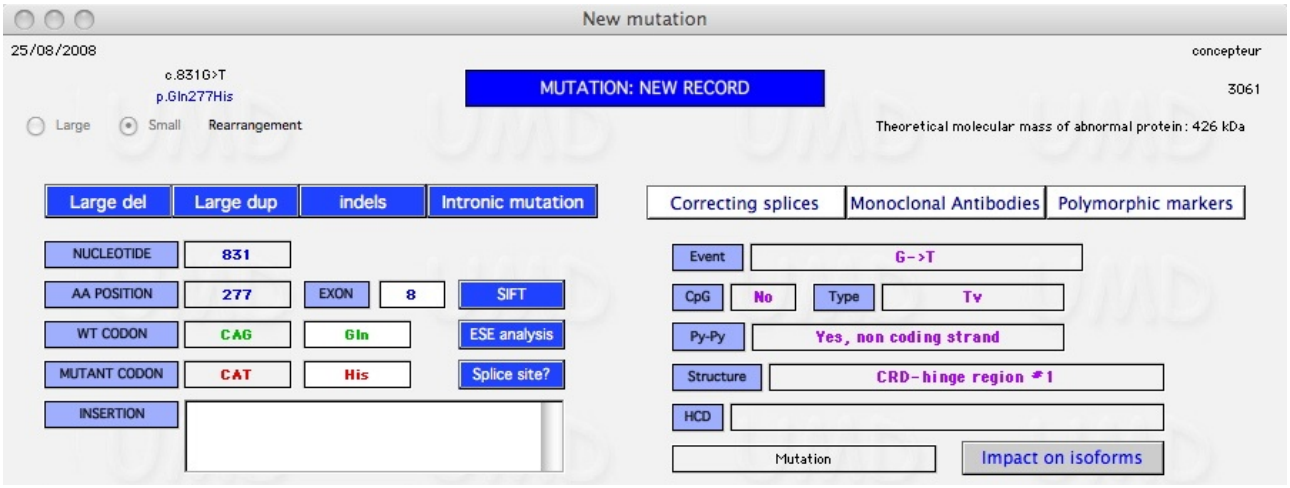
1) c.10018T>C (p.Cys3340Arg)

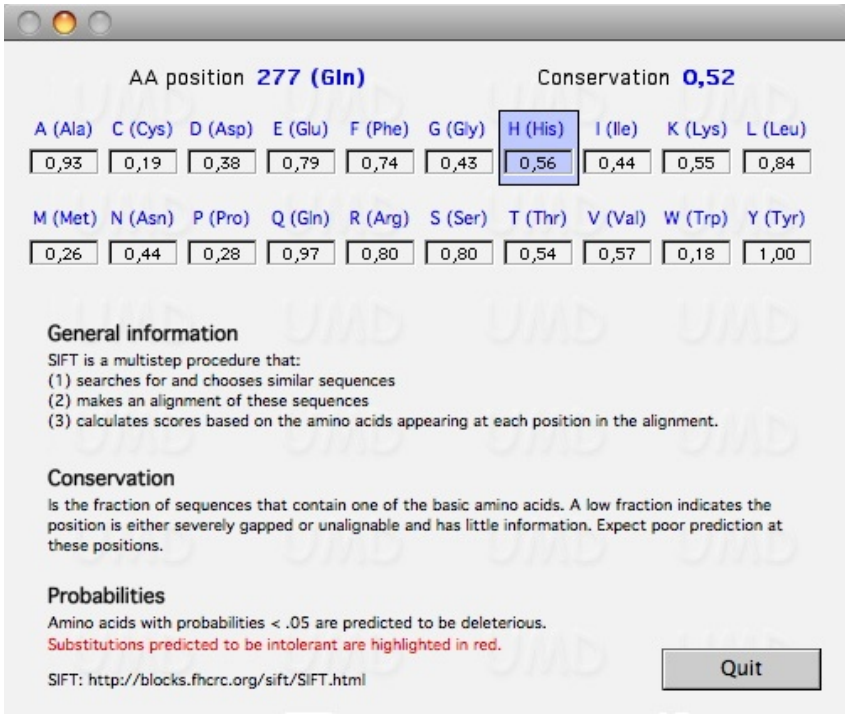


A click on the SIFT button, allow you to display SIFT predictions for this specific position:

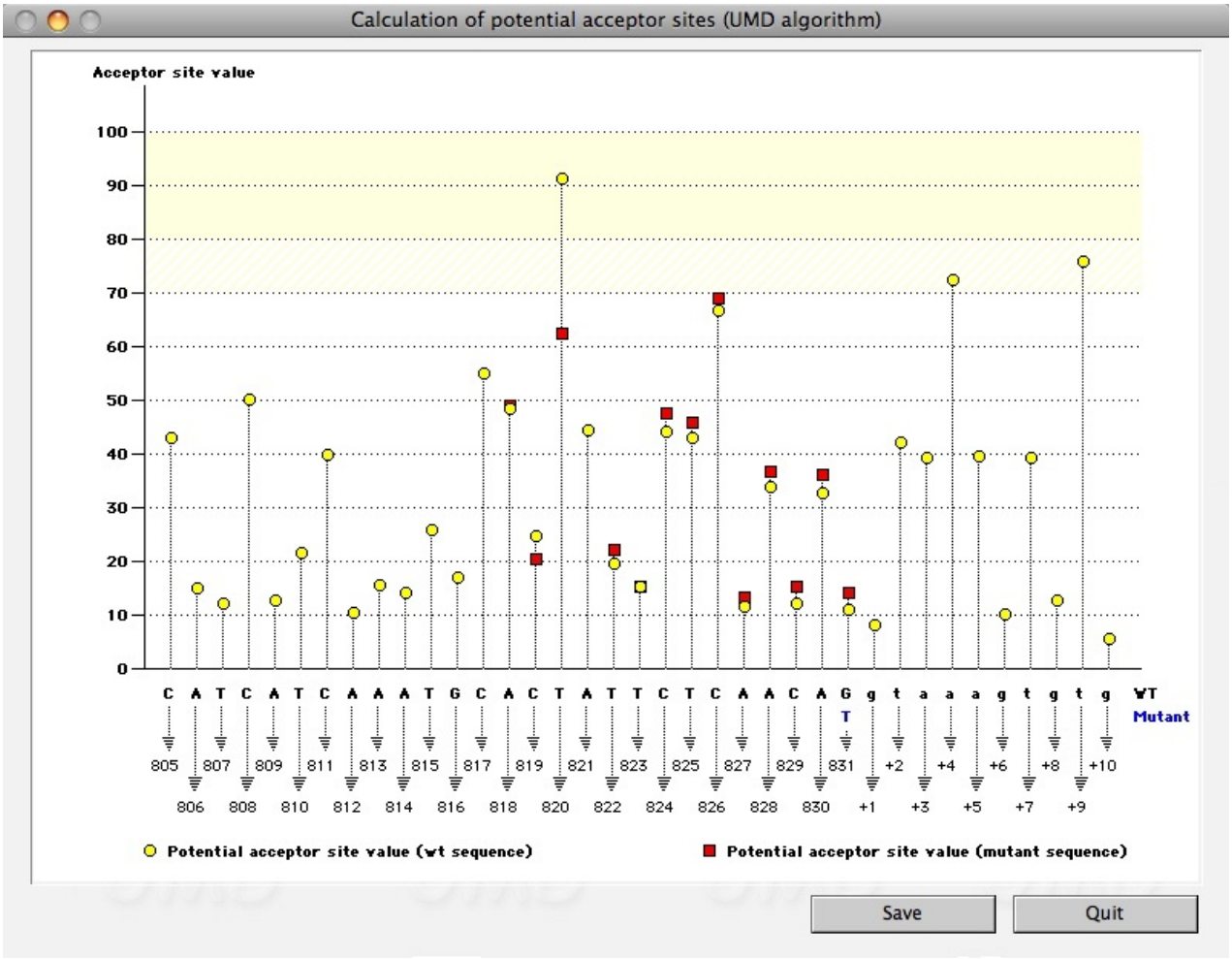


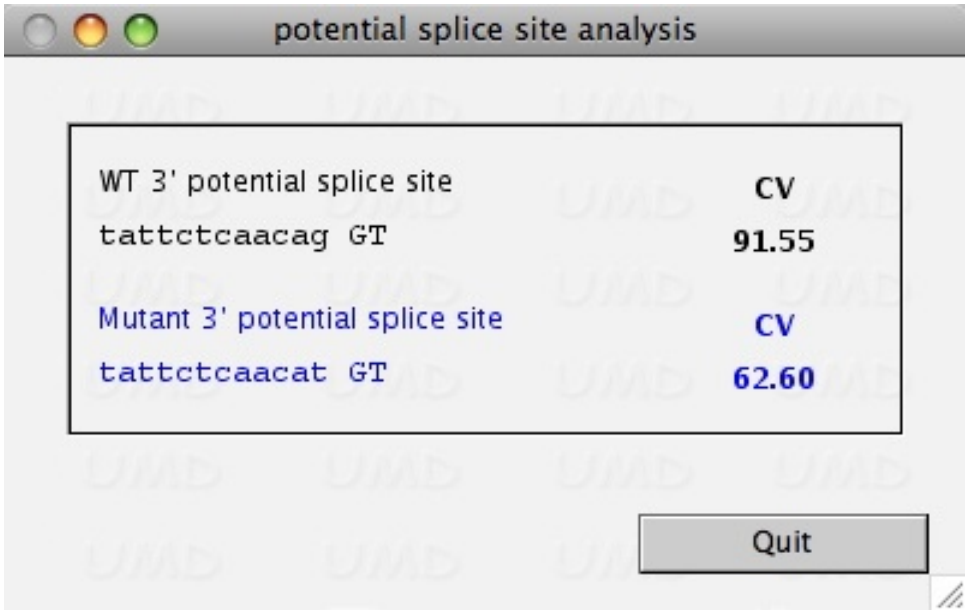
2) c.831G>T (p.Gln277His)





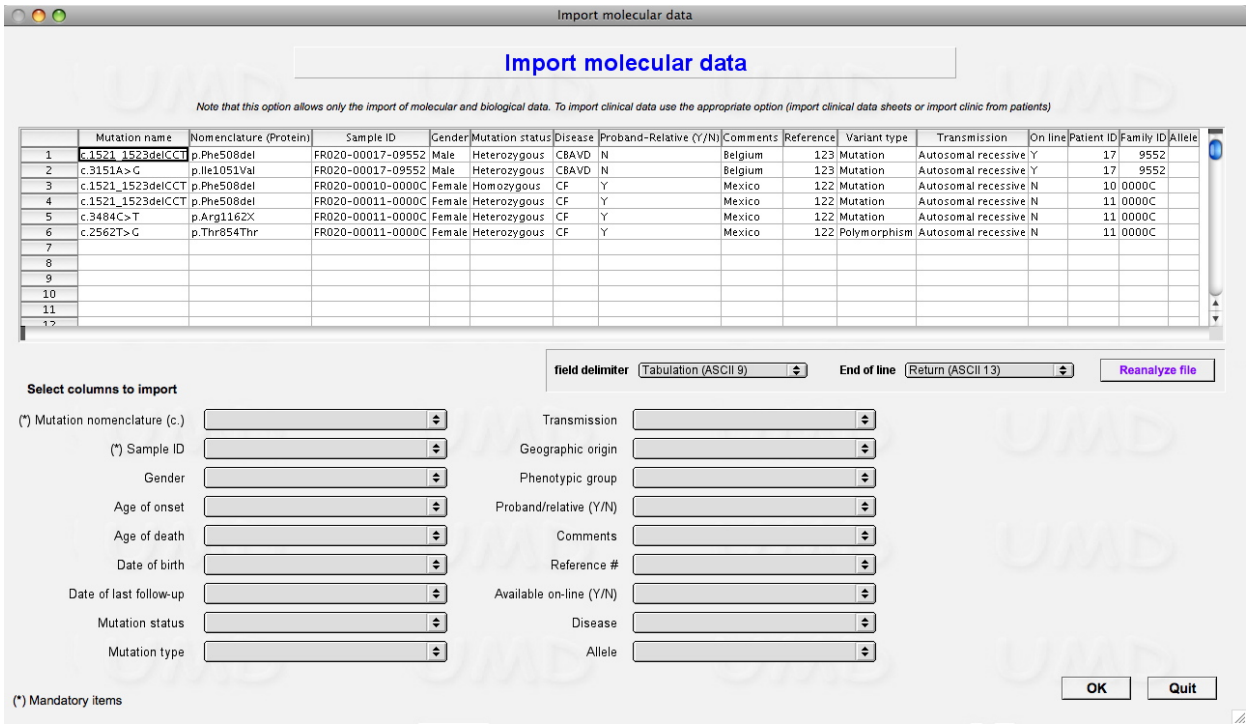
Note that SIFT does not predict this mutation as being pathogenous while UMD Predictor® predicts that it is indeed a pathogenous mutation as it involves the last nucleotide of the exon and therefore disrupt the donor splice site. To access this level of information, click on the "***Splice site?***" button. This will give you access to the same graphical display used for intronic mutations.


Calculation of potential acceptor sites (UMD algorithm)
Acceptor site value
100
90
80
70
60
50
40
30
20
10
0
C A T C A T C A A A T G C A C T A T T C T C A A C A G g t a a a g t g t g WT
T Mutant
805 807 809 811 813 815 817 819 821 823 825 827 829 831 +2 +4 +6 +8 +10
806 808 810 812 814 816 818 820 822 824 826 828 830 +1 +3 +5 +7 +9
Potential acceptor site value (wt sequence)
Potential acceptor site value (mutant sequence)
Save
Quit

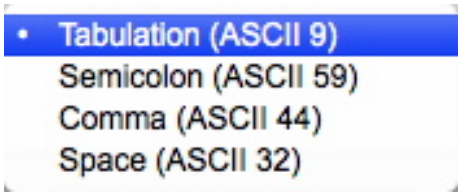

potential splice site analysis
WT 3' potential splice site
CV
tattctcaacag GT
91.55
Mutant 3' potential splice site
CV
tattctcaacat GT
62.60
Quit

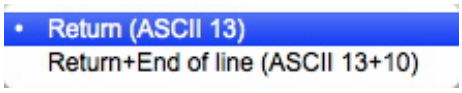
IV- IMPORT MOLECULAR DATA INTO UMD®

When setting up a new UMD-LSDB, the curator has sometimes data already stored into a tab-delimited file format. In this situation, he can import these data directly into UMD® using the "***Import data***" option from the "***File***" menu.

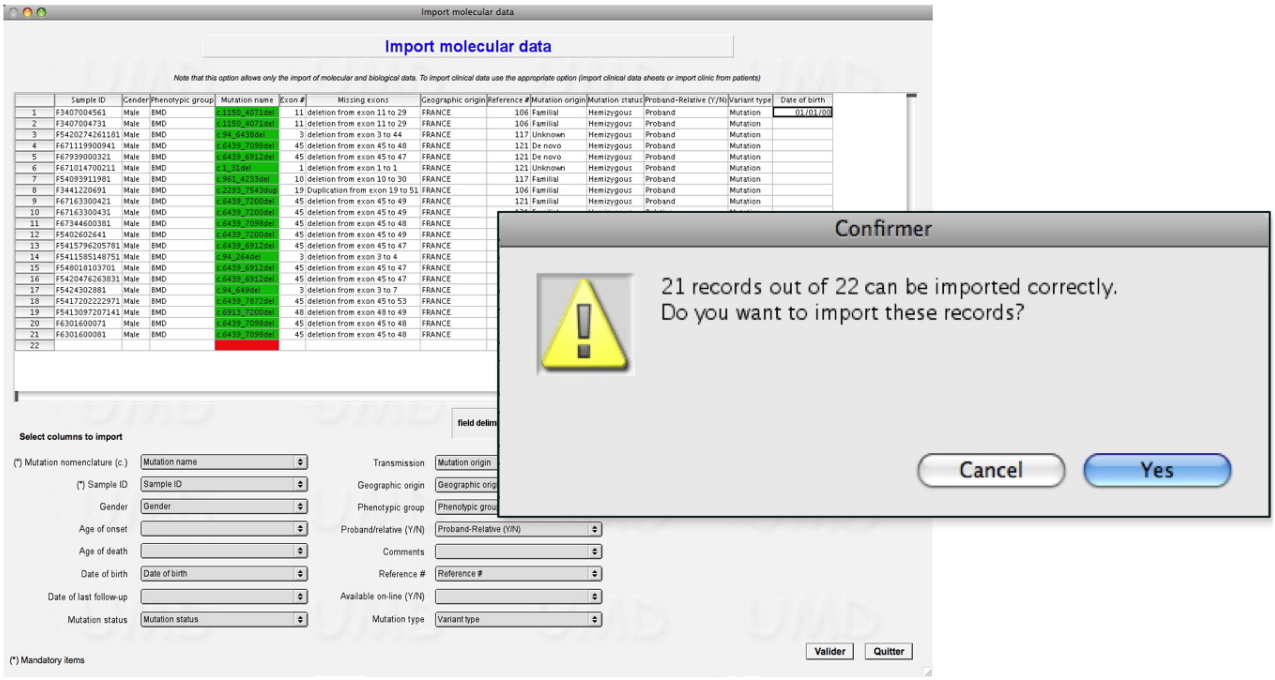


This option allows the user to select a data file and to display its content into UMD®:
The content of the original record is displayed as a table (top). If the file is not displayed correctly, the user can define different parameters for field delimiters and end of line.

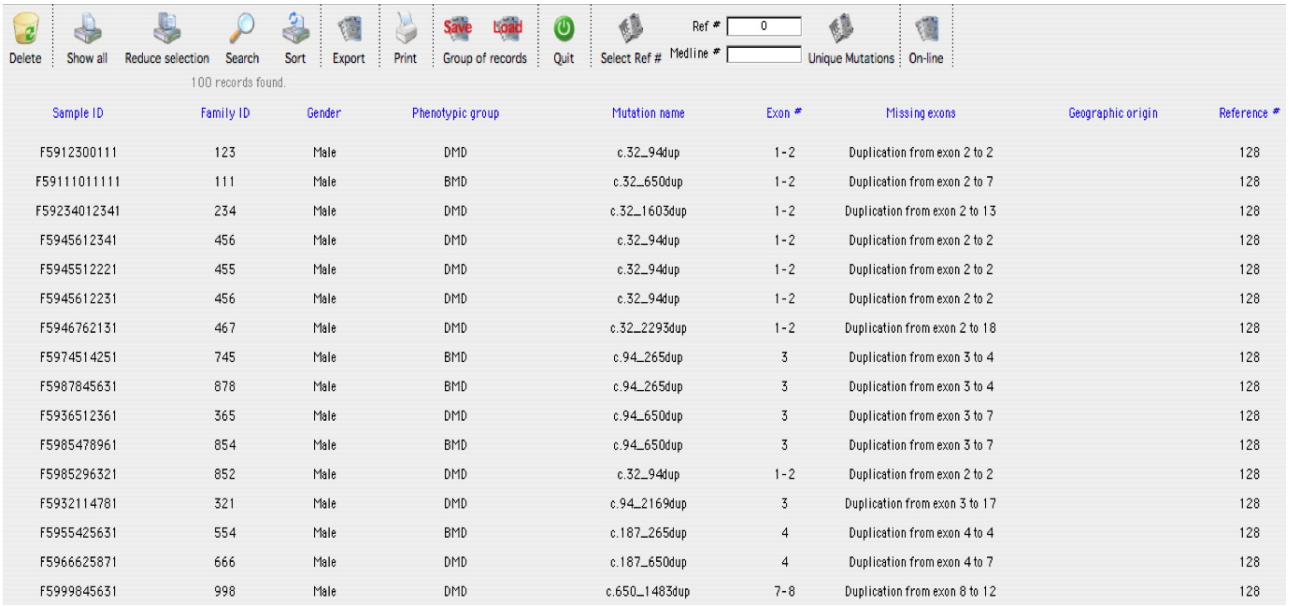




When the parameters have ben correctly set-up, press the "***Reanalyze file***" button.
If the file content is correctly displayed, you can now use the bottom part of the screen to select the various columns to import and define field correspondence. Two fields are mandatory "***Mutation nomenclature (c.)***" and "***Sample ID***". The other fields are optional but highly recommended. For each field, a pop-up menu allows you to select one of the columns from your file. If you select multiple times the same column, only the last selection will remain active all previous selections will be removed.
After the column selection process, press the "***OK***" button. The UMD® software will then check if the international nomenclature name is correct (mutation name will appear in green) or not (mutation name will appear in green). At this stage you can either confirm the import of modify the table content.



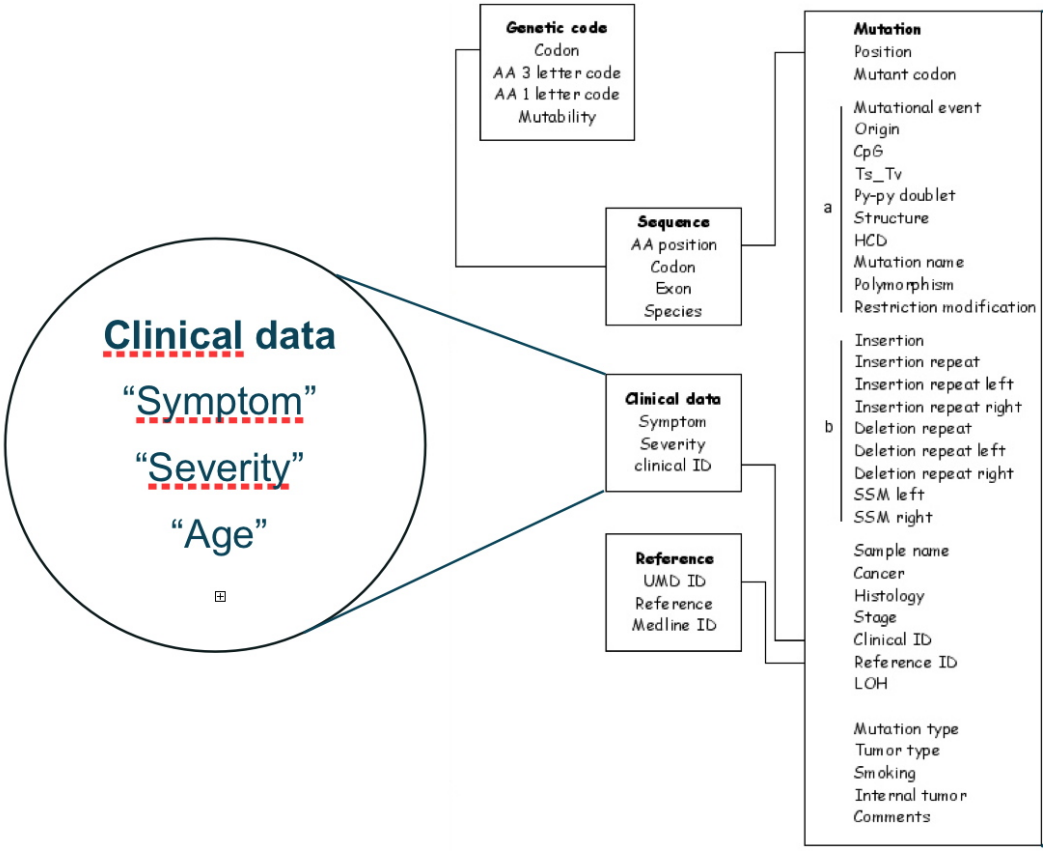
After the import process, mutations can be checked using the "***View-Modify mutations***" option from the "***Show all mut.***" menu.



Sample ID	Family ID	Gender	Phenotypic group	Mutation name	Exon #	Missing exons	Geographic origin	Reference #
F5912300111	123	Male	DMD	c.32_94dup	1-2	Duplication from exon 2 to 2		128
F59111011111	111	Male	BMD	c.32_650dup	1-2	Duplication from exon 2 to 7		128
F59234012341	234	Male	DMD	c.32_1603dup	1-2	Duplication from exon 2 to 13		128
F5945612341	456	Male	DMD	c.32_94dup	1-2	Duplication from exon 2 to 2		128
F5945512221	455	Male	DMD	c.32_94dup	1-2	Duplication from exon 2 to 2		128
F5945612231	456	Male	DMD	c.32_94dup	1-2	Duplication from exon 2 to 2		128
F5946762131	467	Male	DMD	c.32_2293dup	1-2	Duplication from exon 2 to 18		128
F5974514251	745	Male	BMD	c.94_265dup	3	Duplication from exon 3 to 4		128
F5987845631	878	Male	BMD	c.94_265dup	3	Duplication from exon 3 to 4		128
F5936512361	365	Male	DMD	c.94_650dup	3	Duplication from exon 3 to 7		128
F5985478961	854	Male	BMD	c.94_650dup	3	Duplication from exon 3 to 7		128
F5985296321	852	Male	DMD	c.32_94dup	1-2	Duplication from exon 2 to 2		128
F5932114781	321	Male	DMD	c.94_2169dup	3	Duplication from exon 3 to 17		128
F5955425631	554	Male	BMD	c.187_265dup	4	Duplication from exon 4 to 4		128
F5966625871	666	Male	DMD	c.187_650dup	4	Duplication from exon 4 to 7		128
F5999845631	998	Male	DMD	c.650_1483dup	7-8	Duplication from exon 8 to 12		128

V- IMPORT CLINICAL DATA INTO UMD®

The structure of the UMD® LSDB is composed of various tables linked by unique fields (relational database). The central table in UMD® is the mutation that is linked to reference sequences, as well as the clinical data as shown below:

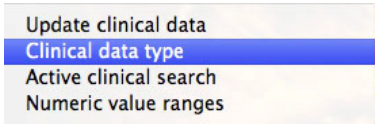


In order to have a very flexible structure that can be adapted to any data set, we designed the clinic table with 3 fields: "***Symptom***", "***Severity***" and "***Age***". The two first are alphanumeric fields that can handle text or numeric values while the third includes only numeric values. As a virtually unlimited number of records can be associated to a single sample (mutation), the user can easily add new clinical descriptions without any modification of the UMD® structure.

On possible drawback of this approach is that it is harder to query numeric values that are included in an alphanumeric field. To circumvent this we have design specific parameters available through the "***File***" menu:

- Clinical data type
- Numeric value ranges

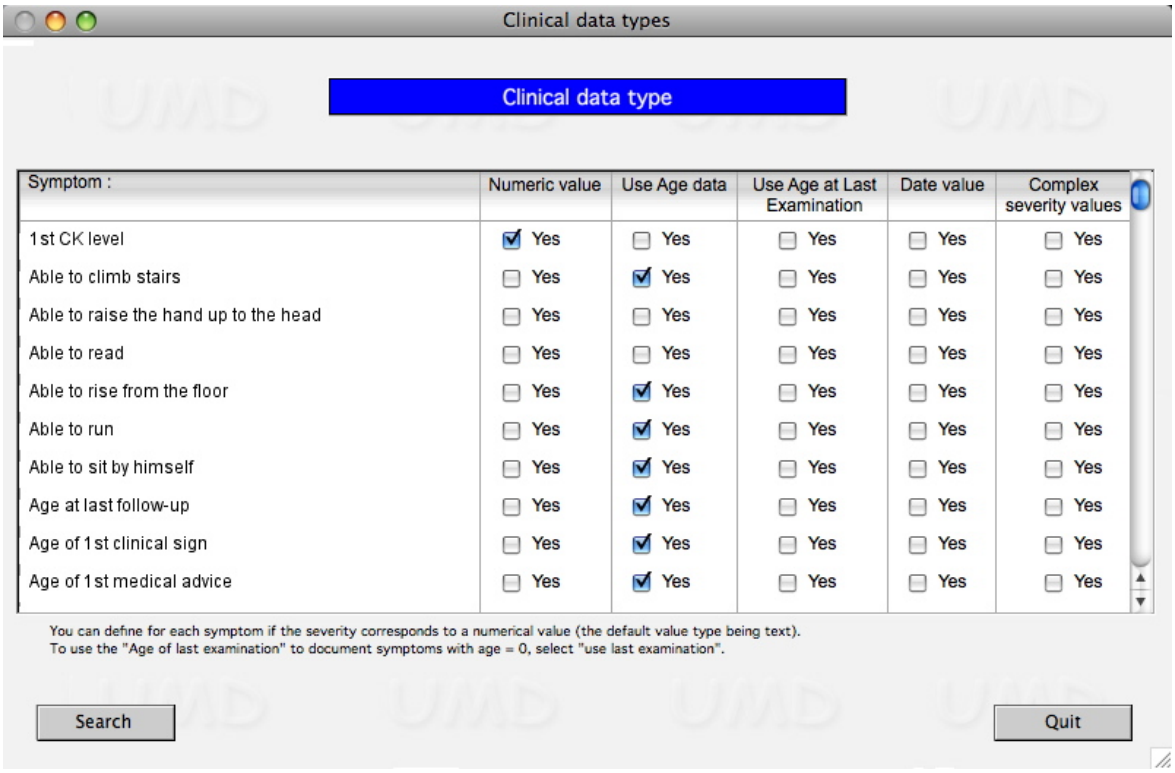
Clinical data type



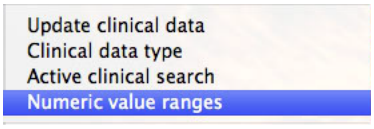
This option allows the user to define for each symptom if the associated severities are either numerical values, dates or simple text. In addition, the user can choose to use the "Age" field or the "Age at last examination" symptom. In fact many users created an "Age at last examination" symptom in order to record when was evaluated for the last time the patient. Finally, when complex answers are entered in the severity field, the user can specify it to the software "Complex severity values". This is particularly useful for symptoms such as "Initial symptom(s)". In this situation the user can enter multiple answers. To do so he should follow a simple rule:

- Each category is separated by a coma
- Sub categories are included in parenthesis and separated by comas

For example: "Motor abnormalities (walk, weakness), Toe walking" means that the patient presented with 2 motor abnormalities (walk and weakness) and toe walking.

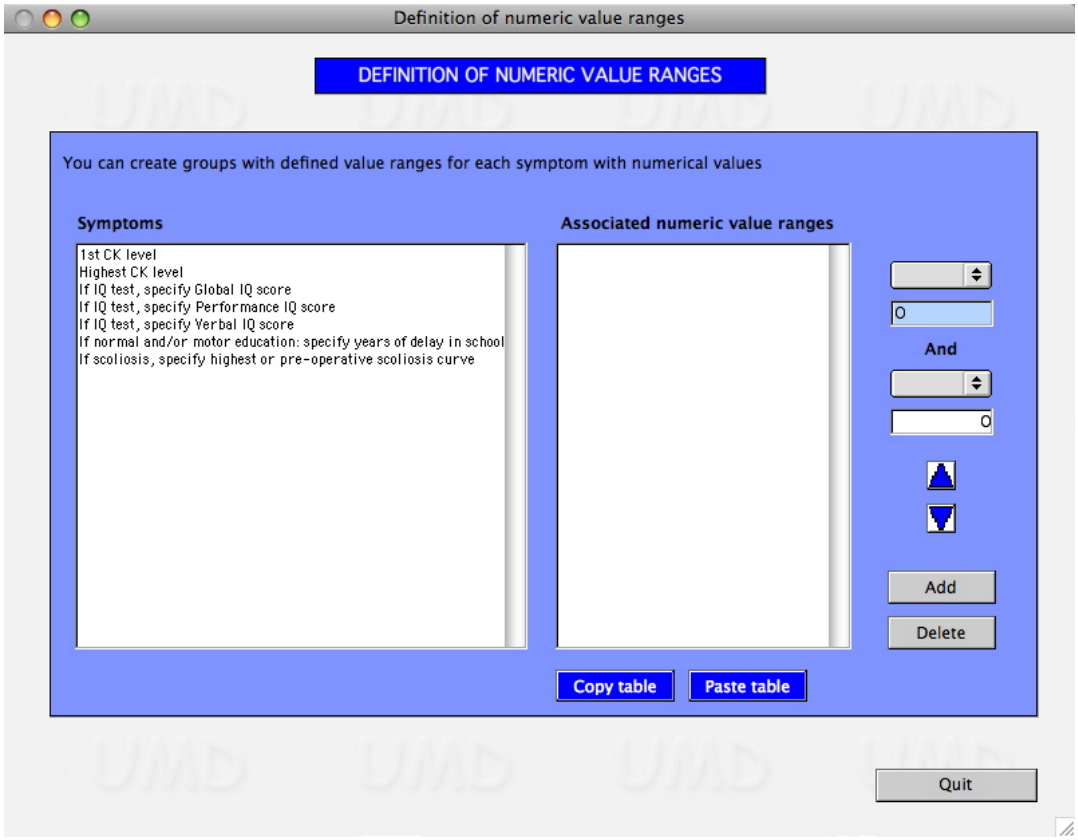


Numeric value ranges



When using numerical values, it is often necessary to define groups in order to analyze data. The "***Numeric value ranges***" option from the "***File***" menu allows to define these groups for symptoms for which the user has defined as numerical values (1st CK level in the previous picture).

The user has access to the following display:



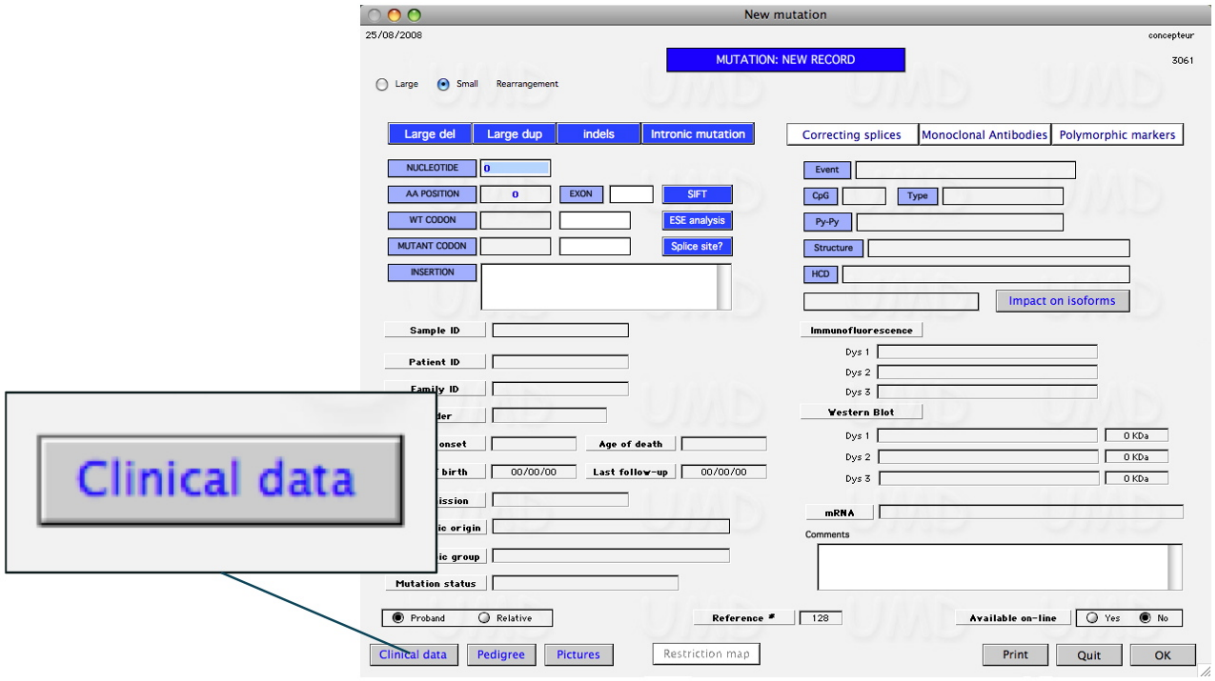
Click on the symptom of interest (left) and define the various groups with the "***Add***" button. For each group define the limits with the 2 pop-up menus (right).
If you want to use the same groups for various symptoms, just use the "***Copy table***" and "***Paste table***" buttons.
The various groups will be available for statistical analysis (see below).

V-a- Various options to import clinical data into UMD®

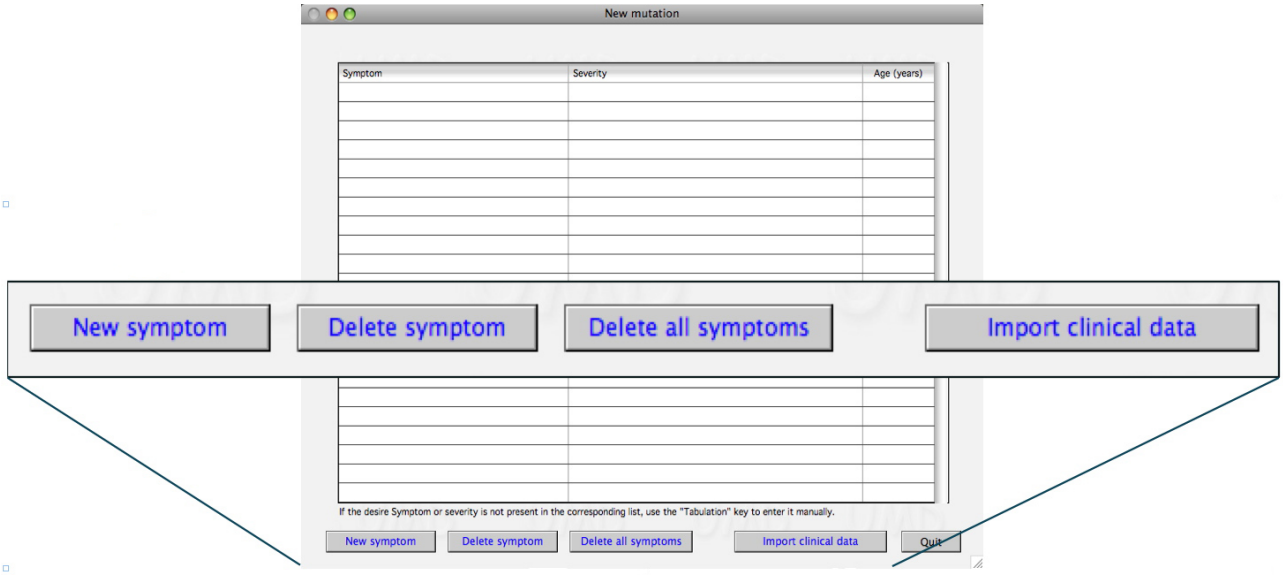
Clinical data can be imported into the UMD® software using four different options:

- Manually during the mutation's record creation
- Via a file (automatic import) during the mutation's record creation
- Via multiple individual files
- Via a single file containing data from multiple patients

1) Manual input during mutation's record creation



On the first screen of the mutation input, a "***Clinical data***" button allows to access to the



clinical data screen.

The user has now access to a simple table containing the three columns: "Symptom", "Severity" and "Age". Four buttons can be used to create "***New symptom***", delete one symptom "***Delete symptom***" or all symptoms "***Delete all symptoms***". A fourth button "***Import clinical data***" can also be used (cf. 2).

When the user clicks on the "***New symptom***" button, he creates a new line in the table. This can be noticed by the 0 value that appears in the "Age" column. He can now double click in the "Symptom" box to get the "Symptom list" in which he can add, delete, sort symptoms. This list is automatically updated when UMD® is launched and will only contain data present in records. It is therefore not necessary to build a list at first as non-used data will be discarded when the software will be re-started. Just add a new symptom when necessary.

When a symptom has been selected from the list, a click on the severity box will display the "Severity list". This list is updated based on the selected symptom and data already present in records. As for symptoms, it is unnecessary to create a long list of severity as unused choices will be discarded when the software will be re-started.

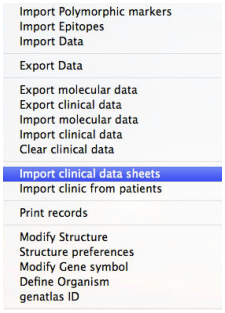
2) Import of clinical data for a single patient

Clinical data are often collected via collaborators. The easiest way is therefore to create ascii files (text tab-delimited files) that can be easily imported during the record creation. To do so, create a table with 2 columns (symptom and severity/age). Each symptom has a label and can include notes, such as the various severity, in parenthesis. These notes will be removed during the import process. An example is given below:

Symptom	Severity [Age]
Phenotype *(DMD, BMD, DMC, IMD, symptomatic carrier, asymptomatic carrier)*	DMD
Age of 1st clinical sign *(Birth /[age] in years /UK)*	[3.5]
Initial symptom *(Motor DD (developmental delay) /Cognitive DD /Motor and cognitive DD /Motor abnormalities (congenital hypotonia, weakness, walk, climbing stairs, running) /Fatigue /Pain: cramping and/or myalgia /Muscle hypertrophy /Toe walking /Cardiomyopathy /UK)*	Motor abnormalities (weakness)
Age of 1st medical advice *(specify [age] in years/UK)*	[3.75]
Diagnostic circumstances *(Clinical sign /Elevated CK /Familial context /Familial context and clinical sign /UK)*	Clinical sign
1st CK level *(in X normal values if available /if no in UI /l /elevated CK /normal CK /UK, specify [age in years])*	
Highest CK level *(in X normal values if available /if no in UI/l /elevated CK /normal CK /UK, specify [age of measurement in years])*	
Sitting acquisition *(Y /Never /UK, if Y: specify [age of acquisition in months])*	Y
Walking alone ? *(Y /Never /UK, if Y: specify [age of acquisition in months])*	Y [24]
Able to run ? *(Y or MFM 30th item = 2-3 or Walton score = 0 /N or MFM 30th item = 0-1 or Walton score ≥ 1 /Never /UK, if N: specify [age of losing the ability])*	Never

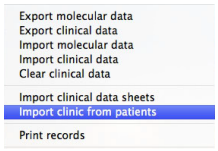
This file structure allows to give directives and help to clinicians and curators during the data collection.
To import such a file into UMD®, use the "***Clinical import data***" button (see upstairs).

3) Global import of individual records



If multiple data are collected from collaborators, it is easier to store each individual data file in the same directory and use the "***Import clinical data sheets***" option from the "***File***" menu.
With this option, the user selects a directory that contains all individual data files. The name of each file should correspond to the sample ID used in the mutation record.

4) Global import of individual records



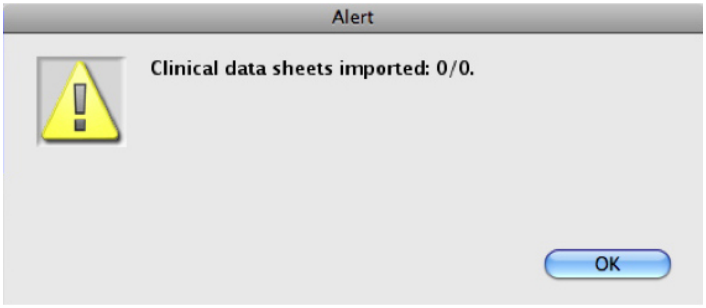
When a single curator is collecting clinical data, he usually likes to store all data in a single file. To do so, he should create a text tab-delimited file that has the following structure:

- The title of each symptom is in the first row
- Tach column contains data for a single symptom
- The first column contains the sample ID
- Each row contain data from a single patient

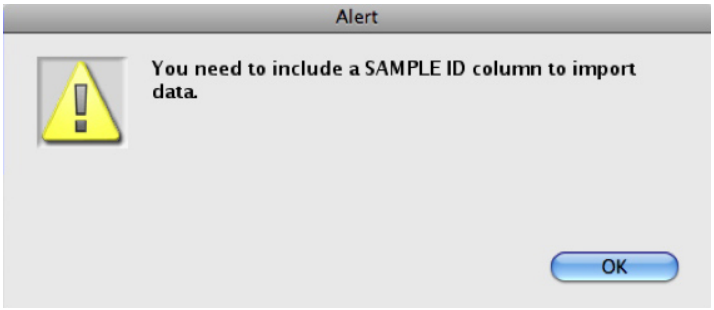
Sample ID	Diagnosis (Phenotype: Duchenne Muscular Dystrophy (DMD) / Becker Muscular Dystrophy (BMD) / Intermediate Muscular Dystrophy (IMD) / Dilated cardiomyopathy (DCM) / symptomatic carrier / asymptomatic carrier / Unknown / I don't know)	Currently able to walk without support (10 steps without another person's help, with or without orthoses: Y or Walton score ≤ 6 or MFM 29th item = 2-3 /N or Walton score ≥ 7 or MFM 29th item = 0-1 /Never /UK, if N: specify [age of losing the ability])	Currently able to sit without support (Able to stand from sitting position without help: Y or MFM 13th item = 2-3 /N or MFM 13th item = 0-1 /Never /UK, if N: specify [age of losing the ability])	Wheelchair use (if over 3 years of age: Y (permanent) / Y (intermittent) / Never / UK, if Y: specify [age])	Scoliosis Surgery (Y /N /UK, if Y: specify [age at surgery])
F754006011	DMD	N [11]	Y	Y (permanent) [8,44]	N
F1300100931	DMD	N [10,33]	Y	Y (permanent) [9]	N
F1300210751	DMD	N [9]	Y	Y (permanent) [7,75]	N
F1300307401	DMD	N [11,25]	Y	Y (permanent) [10,25]	N
F1300410231	IMD	N [14,5]	Y	Y (permanent) [12,25]	N
F1301401051	DMD	N [9]	N [14]	Y (permanent) [10]	Y [14,44]

After selecting the "***Import clinic from patients***" option of the "***File***" menu, the user is prompted to select a text file containing clinical data. The UMD® software will then check for the sample ID and if such a sample exist will import the clinical data. If previous data are available for this sample, they are discarded. This process should therefore be used carefully. At the end of the process, a message will inform the user of the number of correctly imported records:

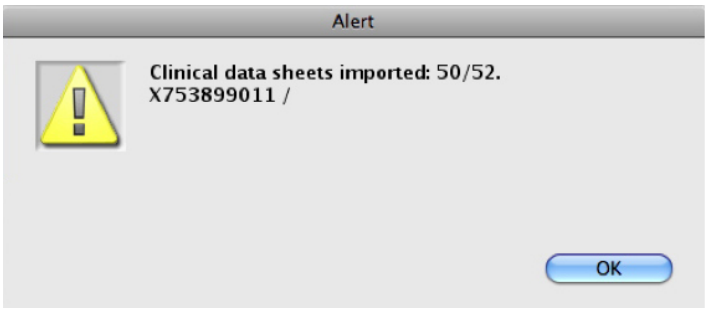
No mutation record has been found with a sample ID corresponding to data from the imported file.



The first column of the imported file is not labelled sample ID and therefore the import process has failed.



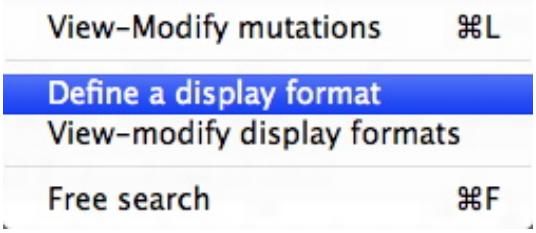
Only 50 out of the 52 rows with clinical data from patients have been successfully imported. The rows corresponding to sample ID "***X753899011***" and "" do not match any mutation record.



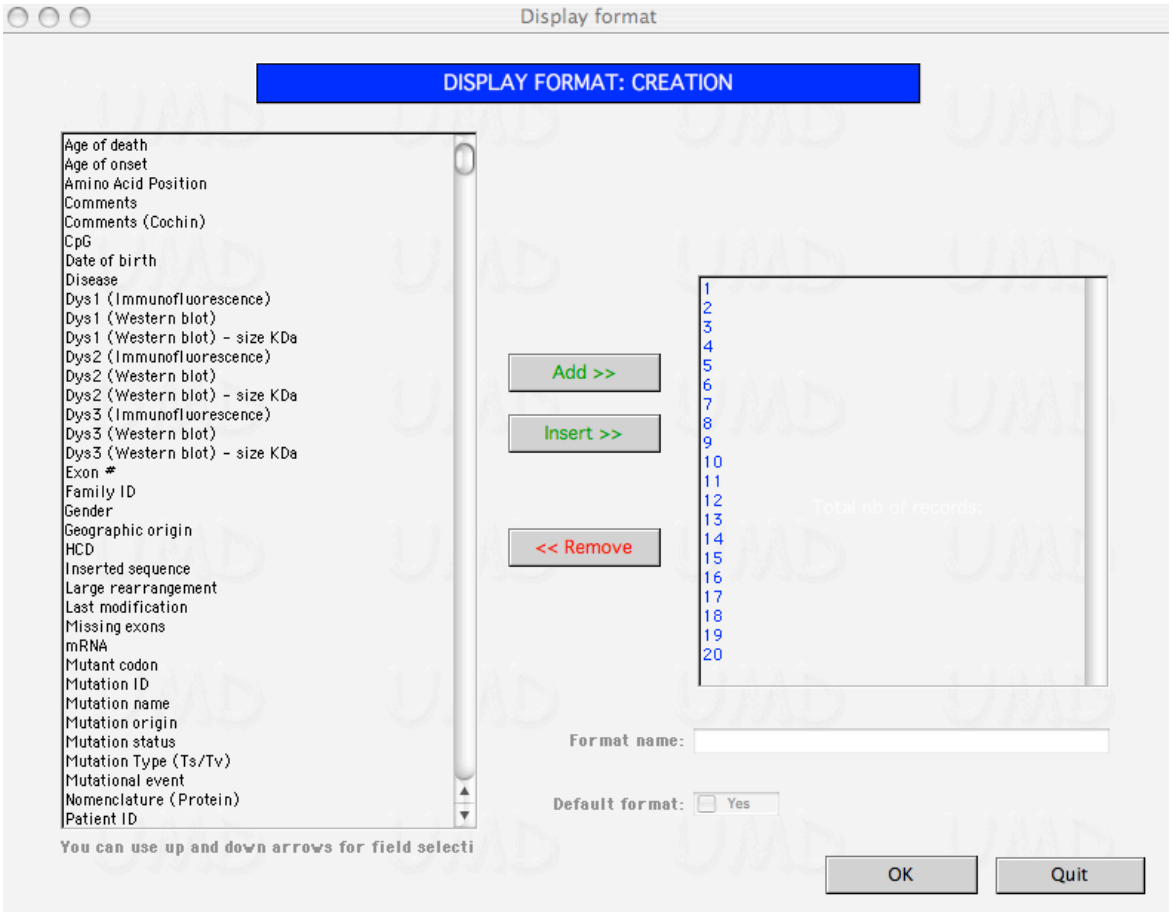
VI DISPLAY FORMATS

VI-a- Creation of display formats

Because all users have different objectives, they would like to visualize data in different formats. Before using the "***View-Modify mutations***" option of the "***Show all mut.***" menu, you then first need to define a display format. To do so, use the "***Define a display format***" option of the "***Show all mut.***" menu.



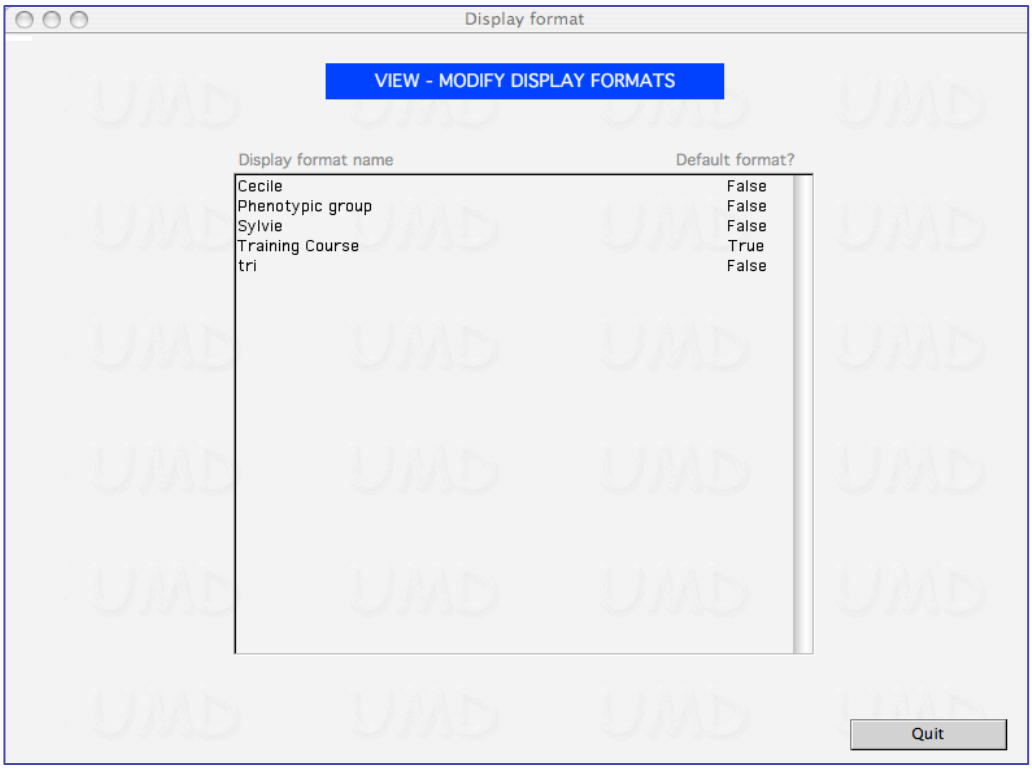
The following screen is then displayed:



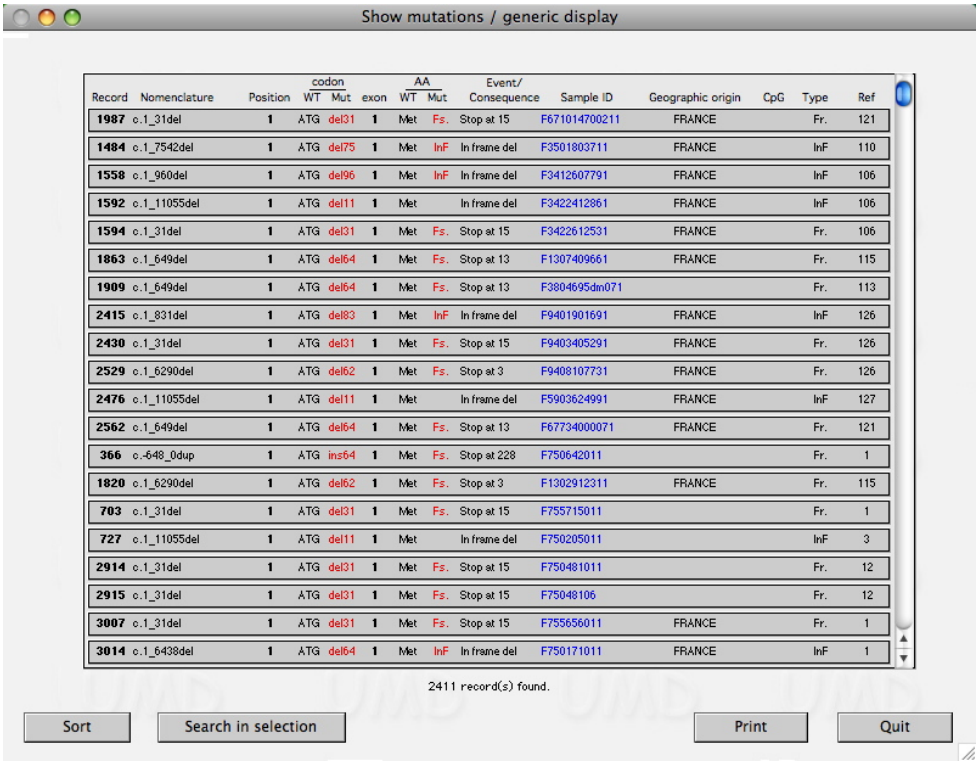
The left table contains the various available fields (except clinical data). You can select up to 20 of these fields to be inserted in a specific display format. Use the "***Add***", "***Insert***" or "***Remove***" buttons to add fields in the right table.
Give a name to this format and decide if it needs to be the default format for display. You can create many formats but only one is active at a time.

VI-b- Modification of display formats

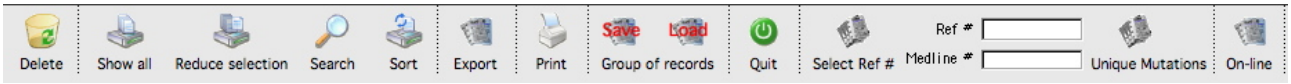
To modify a display format use the "***View-modify display formats***" option from the "***Show all mut.***" menu. A list of all available display formats is available. The default format is labelled "True".



To modify a format, double-click on the corresponding line. This allows a rapid change of the default format or a more complex change of the format itself. Note that if no default format is selected the following generic format is used:



When you use a display format, you create a list that includes selected columns. All formats share a common set of buttons:



Whose functions are:

Delete	To delete selected records and associated data (clinics, pictures).	Print	To print the list of records using the columns from the display format
Show all	To display all records.	Save Load Group of records	To create or load a list of records.
Reduce selection	To reduce the list to selected records.	Quit	To quit the list.
Search	Gives access to the search interface to select mutations using multiple criteria.	Select Ref # Ref # Medline #	To select records associated to a specific reference.

 Sort	To sort records using one or multiple criteria.	 Unique Mutations	To search unique mutations associated to a selected reference.
 Export	To export data limited to the columns from the display format.	 On-line	To set the status of all records from the list to "available on-line".

Note that you can select multiple lines using the "Command" (Mac) or "Control" (PC) buttons. This allows you to either reduce the selection to these selected records or to delete them.

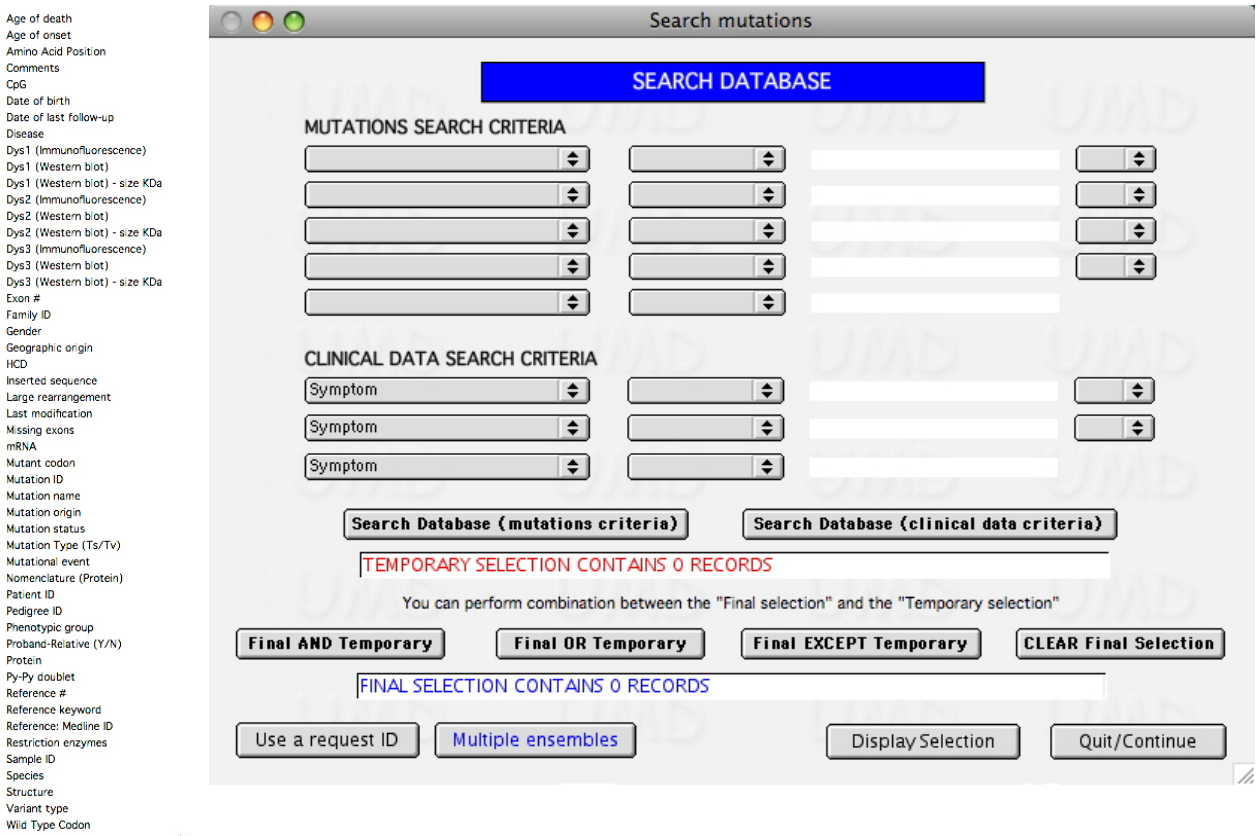
VII SEARCH INTERFACES

In order to facilitate the selection of records, two search interfaces can be used within the UMD® software: the "***UMD-web search interface***" and the "***Regular search interface***".

VII-a- UMD-web search interface

This interface gives access to a limited list of fields from the "*Mutation*" table and to the 3 fields from the "*Clinical data*" table. The user can combine criteria and perform multiple searches. The use of the "*Search Database (mutations criteria)*" button use the fields from the "*Mutation*" table while the "*Search Database (clinical data criteria)*" button use the fields from the "*Clinical data*" table. The number of records corresponding to the request is display as a "Temporary selection". If the user wants to use this set of records, he can transfer the temporary records into the "*Final selection*".

The list of available fields from the "*Mutation*" table is listed below:



The "*Final AND Temporary*", "*Final OR temporary*", "*Final EXCEPT Temporary*" and "*CLEAR Final Selection*" buttons allow a progressive selection of records combining results from the "Final" and the "Temporary" selections.

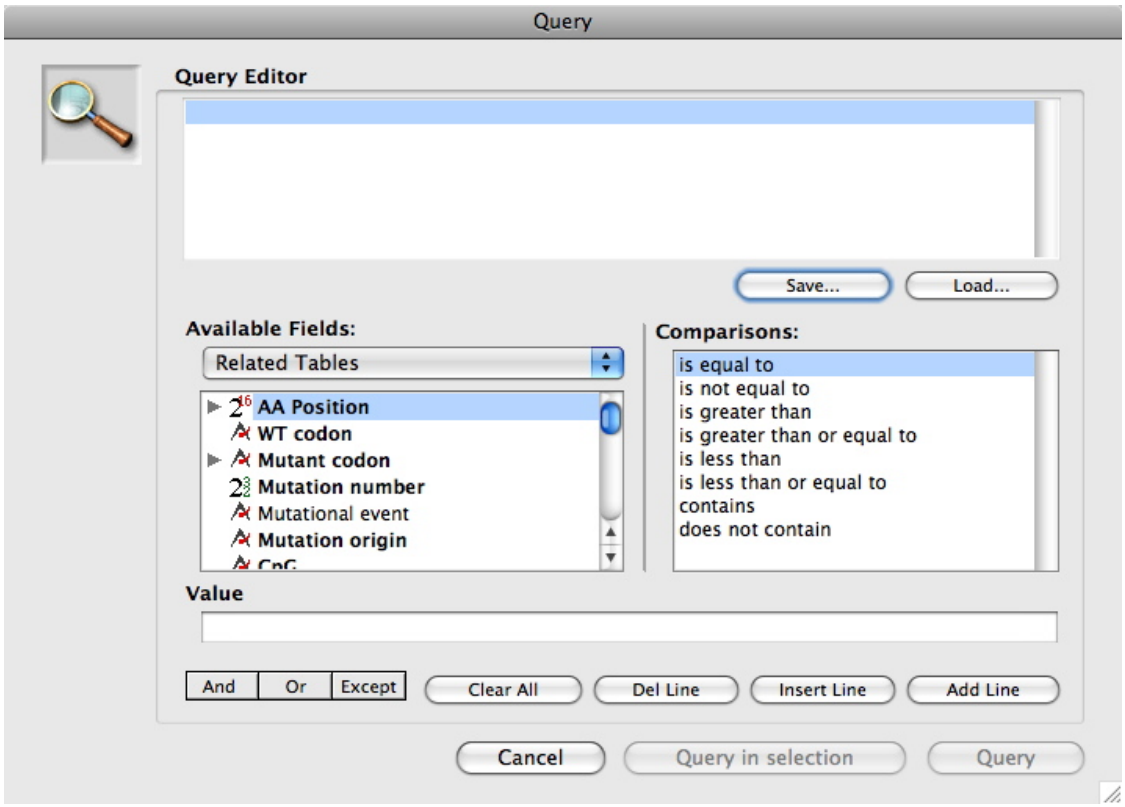
"*Final AND Temporary*": Records should belong to the "Final" and the "Temporary" group of records to be selected;

"*Final OR temporary*": Records should belong to either the "Final" or the "Temporary" group of records to be selected;

"*Final EXCEPT Temporary*": Records should belong to the "Final" but not to the "Temporary" group of records to be selected;

VII-b- Regular search interface

This interface gives access to all fields from the "*Mutation*" and related tables from the UMD database. Note that all fields are available even the ones that are not relevant for your

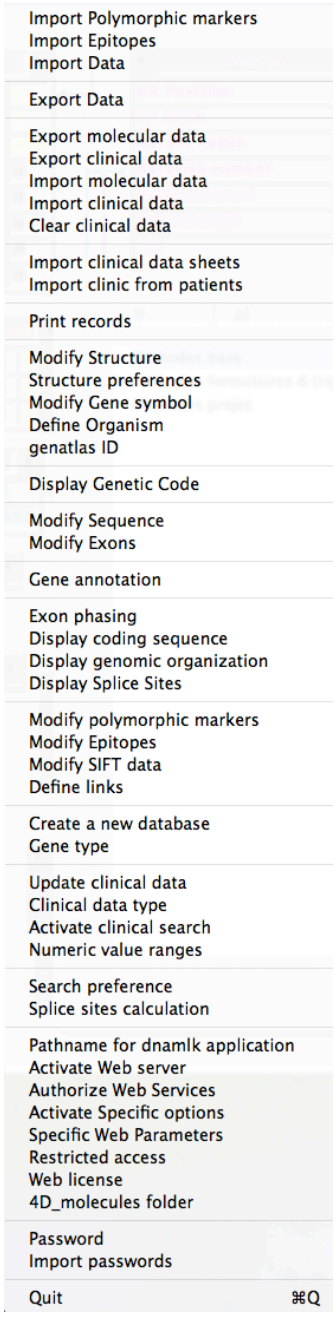


database. We therefore recommend that you use the "***UMD-web search interface***" instead.

If you are familiar with the 4th Dimension software you may choose to use this regular interface. Note that you can add as many queries as necessary in the Query box using the "*Add Line*", "*Insert Line*" and "*Del Line*" button. You can combine queries using the "*And*", "*Or*" and "*Except*" buttons.
In addition, you can "Save" and "Load" your queries with the corresponding buttons.

VIII DESCRIPTION OF ALL FUNCTIONS

All functions of the UMD® software are available through eleven menus: "***File***", "***References***", "***Mutations***", "***Pedigree***", "***Show all mut.***", "***Statview***", "***Phenotype***", "***Genotype***", "***Haplotype***", "***Introns***", "***Modules***" and "***Variations***".



VIII-a- The "File" menu

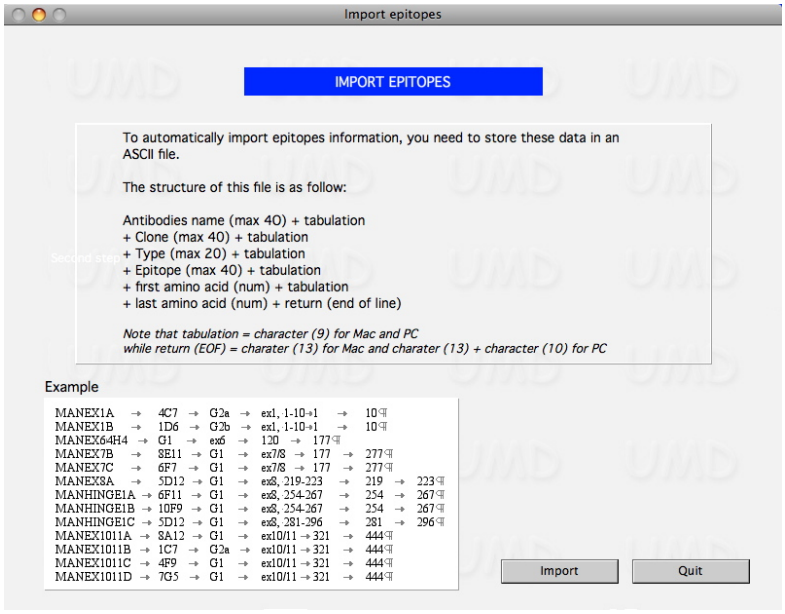
This menu contains 48 functions, which are mainly use to set preferences, import/export data and activate specific tools:

1) Import Polymorphic markers

This function is used to import SNP data (cf. II-b).

2) Import Epitopes

This function is used to import data about antibodies. The file must



contain 6 columns as indicated below:
After the initial import, data can be modified using the "Modify epitopes" function (cf. VIII-a-x).

3) Import Data

This function allows the import of molecular data from any source (cf. IV).

4) Export Data

This function allows the export of all data (molecular and clinical) from a specific set of records freely chosen by the user.Note that clinical data should have been correctly defined within the UMD software by using the "Update clinical data" function (cf. VIII-a-x).

5) Export molecular data

This function is used to combine data from multiple curators. It should be used with the "*Import molecular data*" function.

6) Export clinical data

This function is used to combine data from multiple curators. It should be used with the "*Import clinical data*" function.

7) Import molecular data

This function is used to combine data from multiple curators. It should be used only with a file generated by the "*Export molecular data*" function.

8) Import clinical data

This function is used to combine data from multiple curators. It should be used only with a file generated by the "*Export clinical data*" function.

9) Clear clinical data

This function is used to clear all clinical data from the database.

10) Import clinical data sheets

This function is used to import multiple individual data sheets (cf. V-a-3).

11) Import clinical from patients

This function is used to import clinical data from multiple patients stored in a single file (cf. V-a-4).

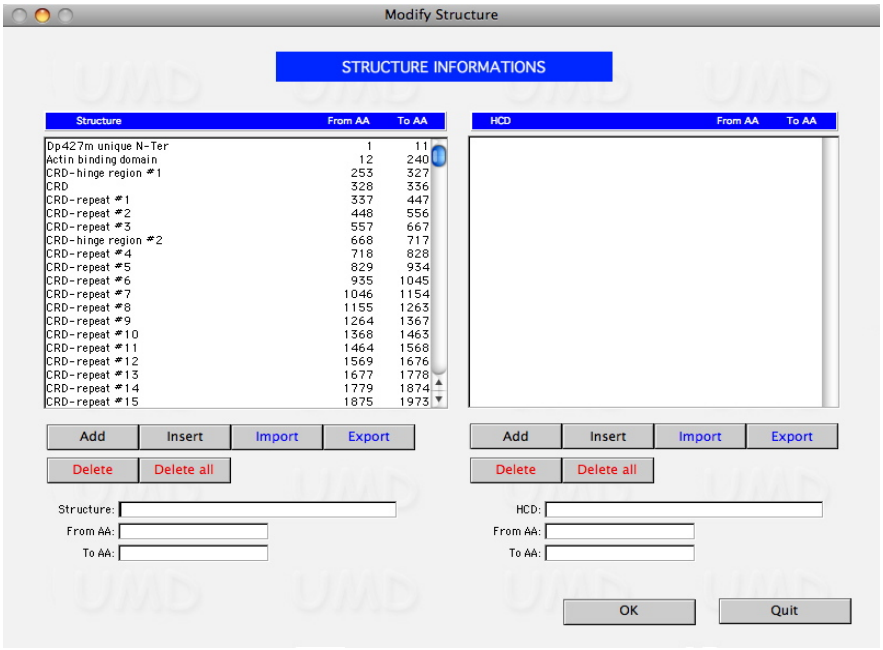
12) Print records

This function is used to print all records form the database. A specific password is required to activate this function. Please contact Christophe Béroud to get this specific password.

13) Modify Structure

This function is used to annotate the structural domains of the protein as well as Highly Conserved Domains (HCD), which represent key residues of the protein. These HCD are used to predict the pathogenicity of missense mutations and therefore a particular attention should be given to their annotation.
The user can either create the various Structural domains and HCD directly into UMD (see below) or import corresponding data from text tab-delimited files containing 3 columns: Structural domain name or HCD name; first amino acid; last amino acid.
Note that structural domains can not overlap and therefore they should be subdivided into unique entities. So a Nuclear localization signal (NLS) located at position 10 to 100 and a

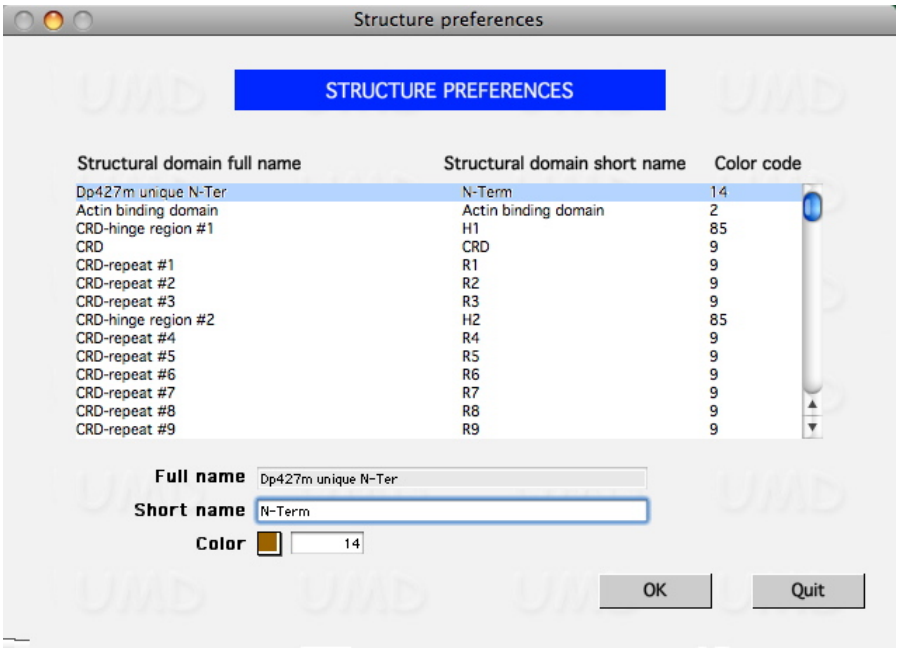
ring domain localized at position 20 to 35 should be annotated: a) NLS 10 to 19, b) NLS + Ring 20 to 35 and c) NLS 36 to 100.



Note that the "*Add*", "*Insert*", "*Delete*" and "*Delete all*" buttons allow the creation or suppression of rows in the corresponding table. After selection of a specific row, the corresponding data can be modified in the fields below the table (Structure/From AA/To AA fields or HCD/From AA/To AA).

14) Structure preferences

This function is used to specify parameters of the structural domains used for graphical display functions.



The user can define a short name for each structural domain and use a color's range of 256 colors.

15) Modify Gene symbol

This function is used to specify the gene symbol during the creation process of a new database (cf. II-a).

16) Define organism

The user can choose between Human and Mouse organism. This will result in different matrices to search for Branch-point sequences.

17) Genatlas ID

This function is used to define the Genatlas ID associated to the Gene. This will be used in a future version of the UMD software to automatically collect annotations.

18) Display Genetic Code

This function is used to display the genetic code using the 3 letters- and the one-letter codes.

19) Modify sequence

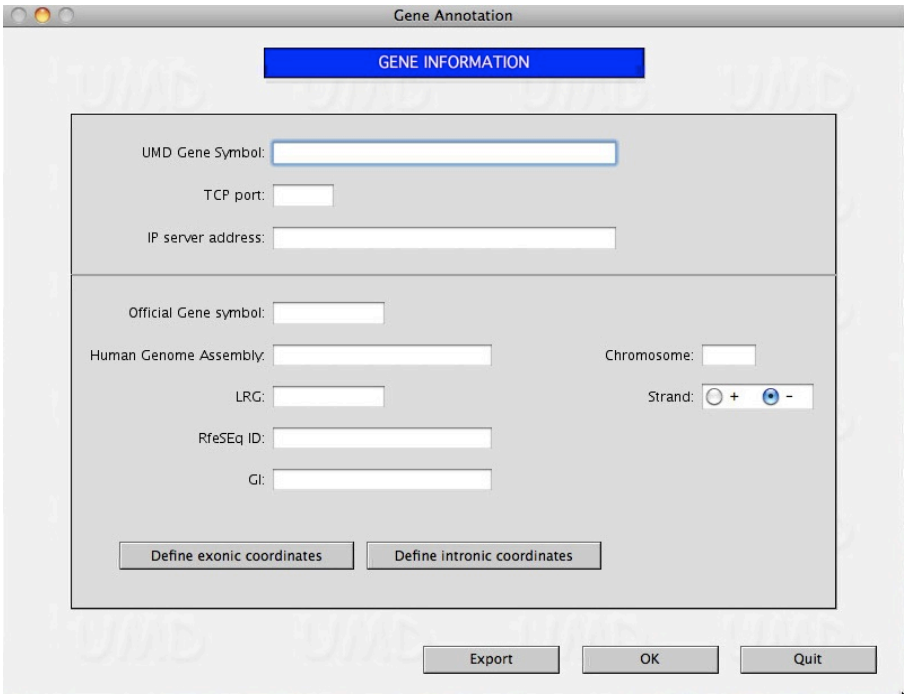
This function is used to display the cDNA reference sequence and eventually to modify it.

20) Modify exons

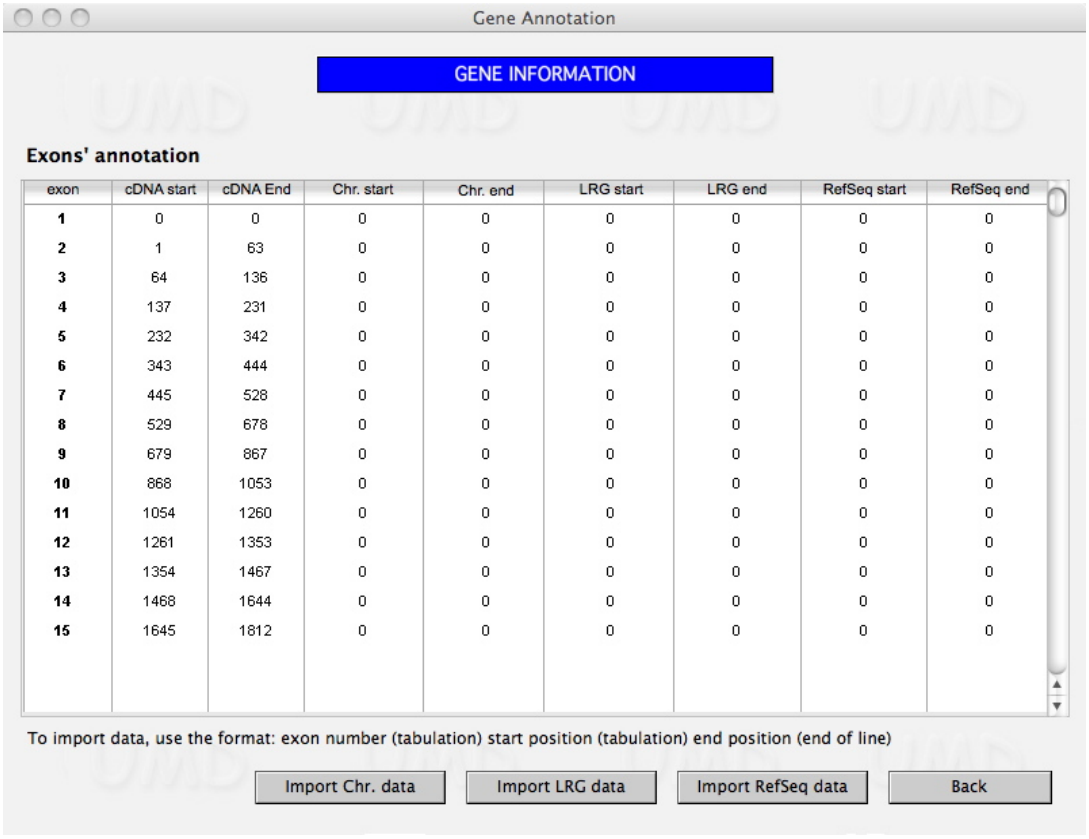
This function is used to display the various exons of the gene and their relative position in the cDNA reference sequence and eventually to modify these data.

21) Gene annotation

This function is used to specify gene information.



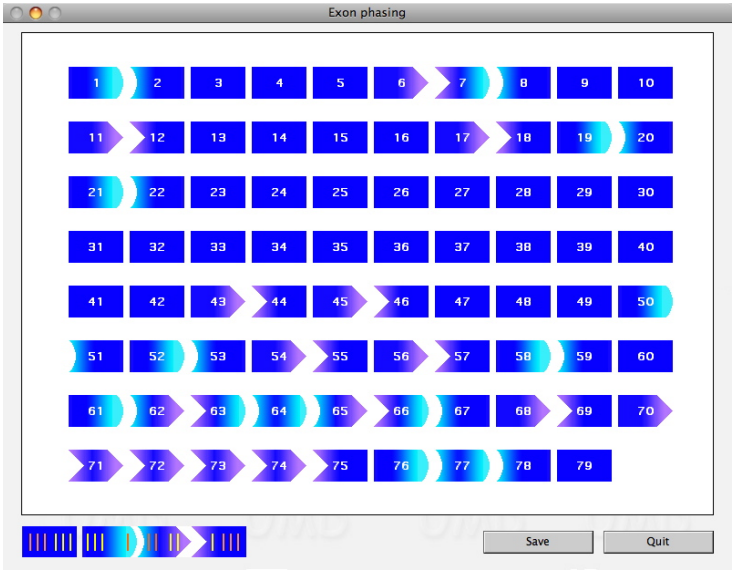
It has been designed to facilitate cross-references of the database using different reference systems such as: The Human Genome Assembly, the LRG and RefSeq systems as well as the GI number from ncbi.
Once the general information is provided, the user can define each coordinate using the "define exonic coordinates" and "define intronic coordinates" buttons.



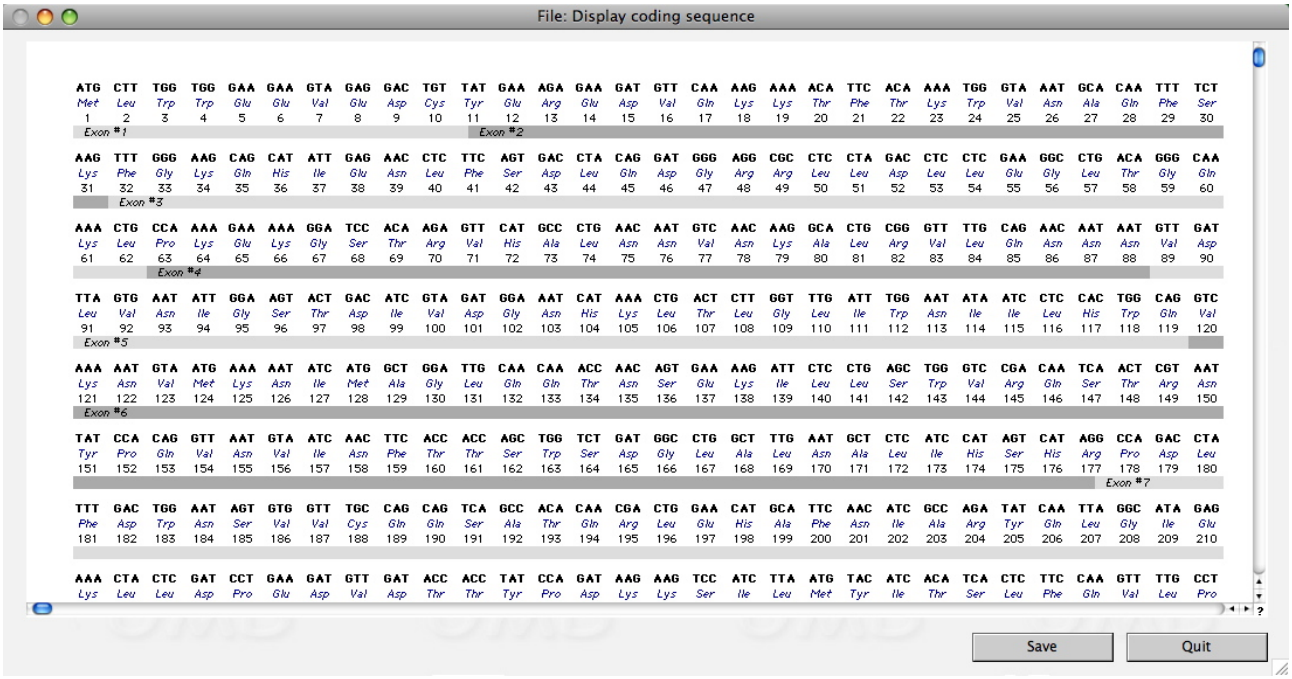
Here is presented the screen linked to the "define exonic coordinates" button. The user can import data from TSV files or manually update each information

22) Exon phasing

Gives access to a graphical display of the various exons. This is particularly useful to evaluate the impact of the deletion of one or more exons on the framing.



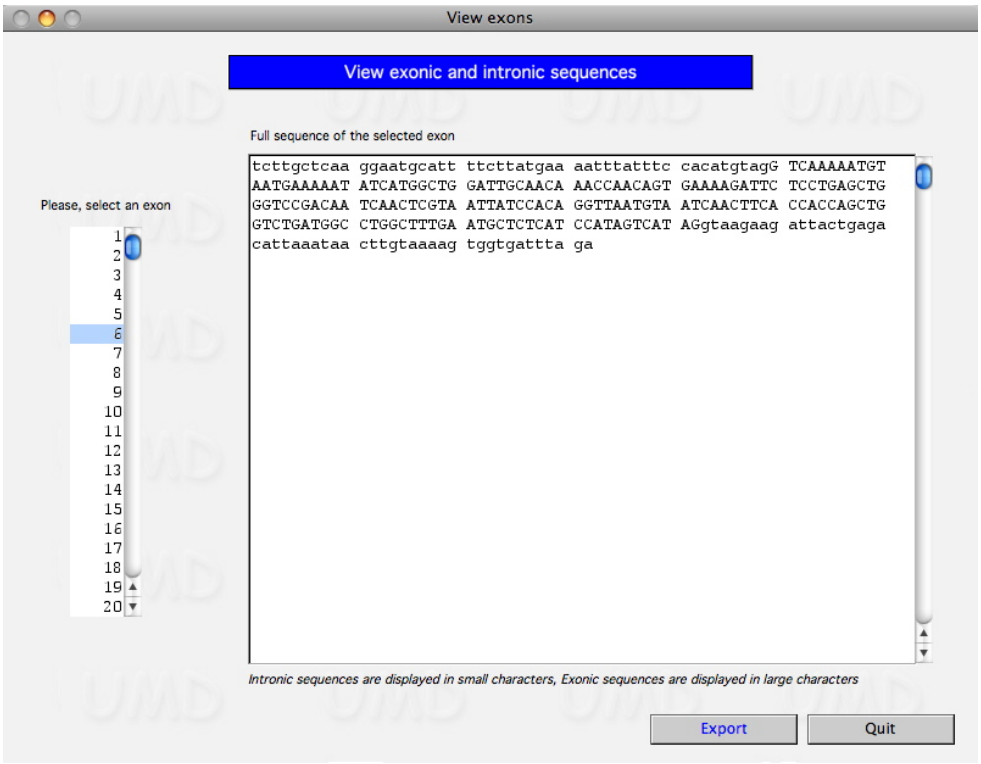
23) Display coding sequence



This function gives access to a graphical presentation of the reference cDNA sequence.

24) Display genomic organization

The user has access to the various exons of the gene with a freely selected portion of the

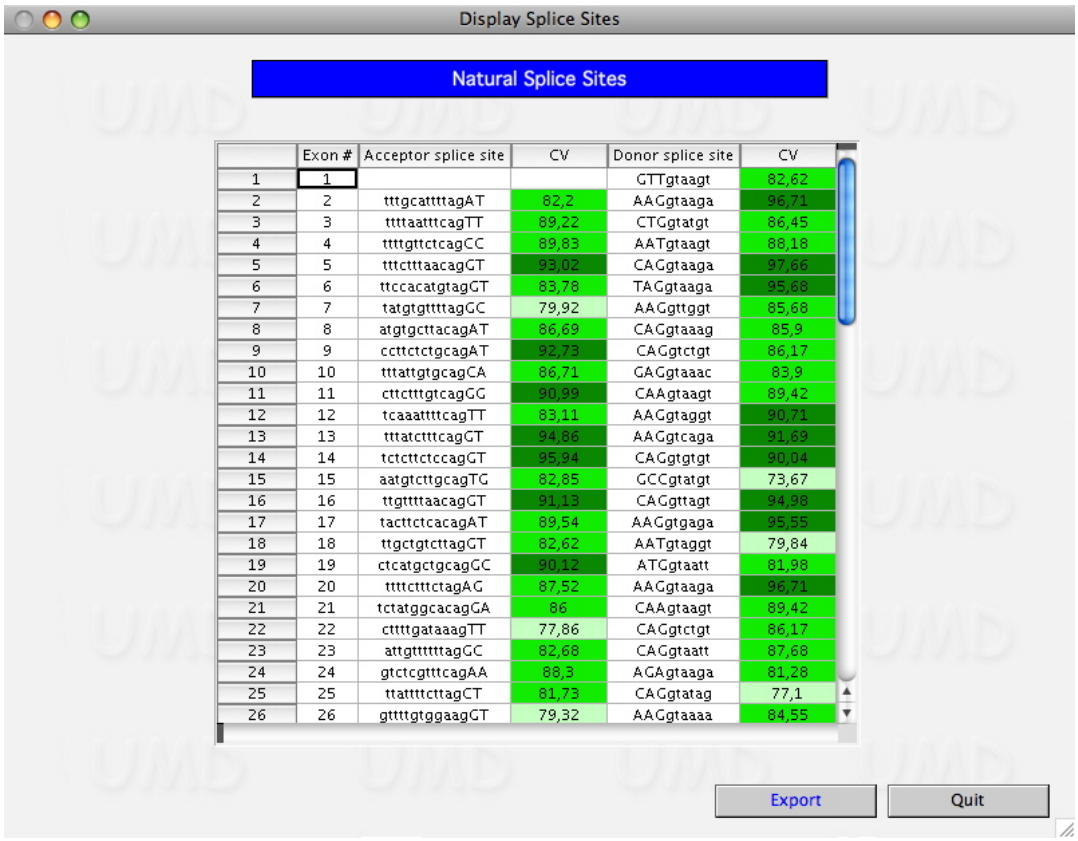


surrounding introns. The corresponding sequences can thus be exported as text file.

Intronic sequences are shown in small letters while exonic sequences are displayed in capital letters.

25) Display splice sites

This function gives access to the natural acceptor and donor splice sites strengths. It could be used to compare the relative strength of cryptic and natural splice sites and facilitate the



	Exon #	Acceptor splice site	CV	Donor splice site	CV
1	1			GTTgtaagt	82,62
2	2	tttgcatttttagAT	82,2	AAGgtaaga	96,71
3	3	ttttaatttcagTT	89,22	CTGgtatgt	86,45
4	4	ttttgttctcagCC	89,83	AATgtaagt	88,18
5	5	tttctttaacagGT	93,02	CAGgtaaga	97,66
6	6	ttccacatgtagGT	83,78	TAGgtaaga	95,68
7	7	tatgtgttttagGC	79,92	AAGgttggt	85,68
8	8	atgtgcttacagAT	86,69	CAGgtaaag	85,9
9	9	ccttctctgcagAT	92,73	CAGgtctgt	86,17
10	10	tttattgtgcagCA	86,71	GAGgtaaac	83,9
11	11	cttctttgtcagGG	90,99	CAAgtaagt	89,42
12	12	tcaaattttcagTT	83,11	AAGgtaggt	90,71
13	13	tttatctttcagGT	94,86	AAGgtcaga	91,69
14	14	tctcttctccagGT	95,94	CAGgtgtgt	90,04
15	15	aatgtcttgcagTG	82,85	GCCgtatgt	73,67
16	16	ttgttttaacagGT	91,13	CAGgttagt	94,98
17	17	tacttctcacagAT	89,54	AAGgtgaga	95,55
18	18	ttgctgtcttagGT	82,62	AATgtaggt	79,84
19	19	ctcatgctgcagGC	90,12	ATGgtaatt	81,98
20	20	ttttctttctagAG	87,52	AAGgtaaga	96,71
21	21	tctatggcacagGA	86	CAAgtaagt	89,42
22	22	cttttgataaagTT	77,86	CAGgtctgt	86,17
23	23	attgtttttagGC	82,68	CAGgtaatt	87,68
24	24	gtctcgtttcagAA	88,3	AGAgtaaga	81,28
25	25	ttattttcttagCT	81,73	CAGgtatag	77,1
26	26	gttttgtggaagGT	79,32	AAGgtaaaa	84,55

interpretation of mutations.
A color code is used to display the strength of splice sites: a) dark green for very strong splice sites; b) green for strong splice sites and c) light green for weaker splice sites. The corresponding sequences and strengths can also be exported in a text file.

26) Modify polymorphic markers

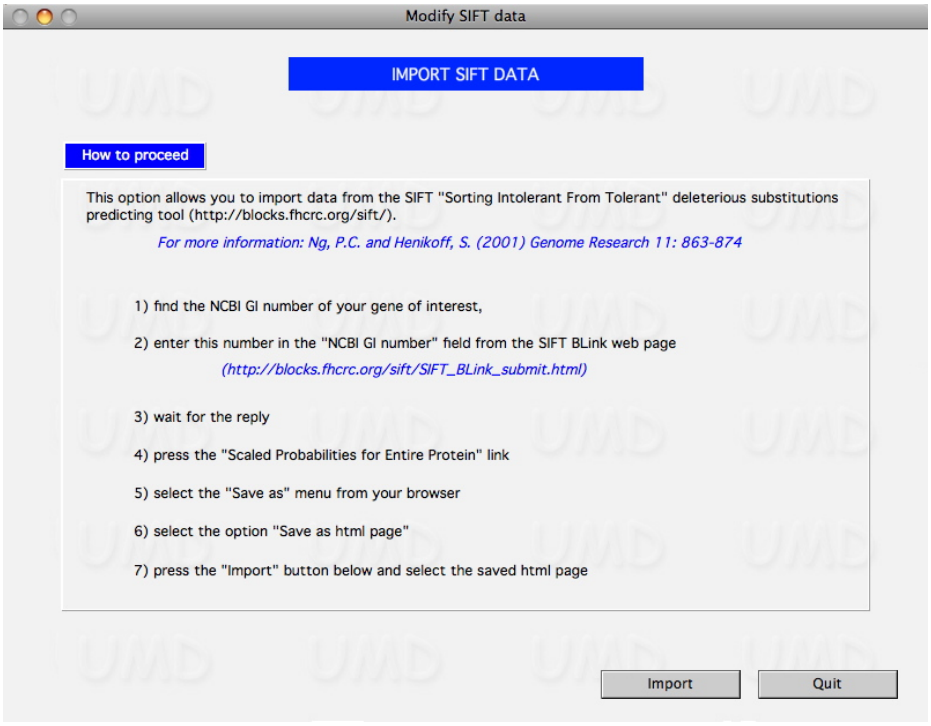
This function is used to modify polymorphic markers or to perform searches from this table.

27) Modify epitopes

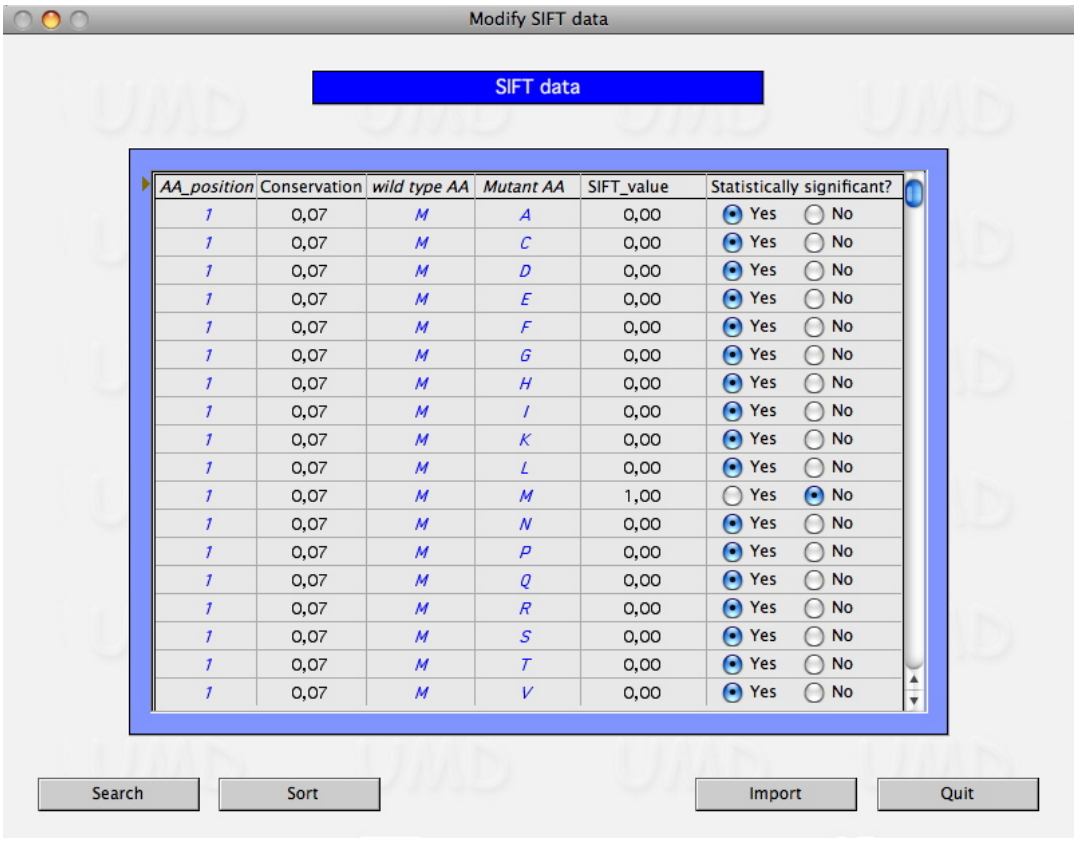
This function is used to modify epitopes or to perform searches from this table.

28) Modify SIFT data

This function gives access to a specific interface to download SIFT predictions. The various steps are available below:



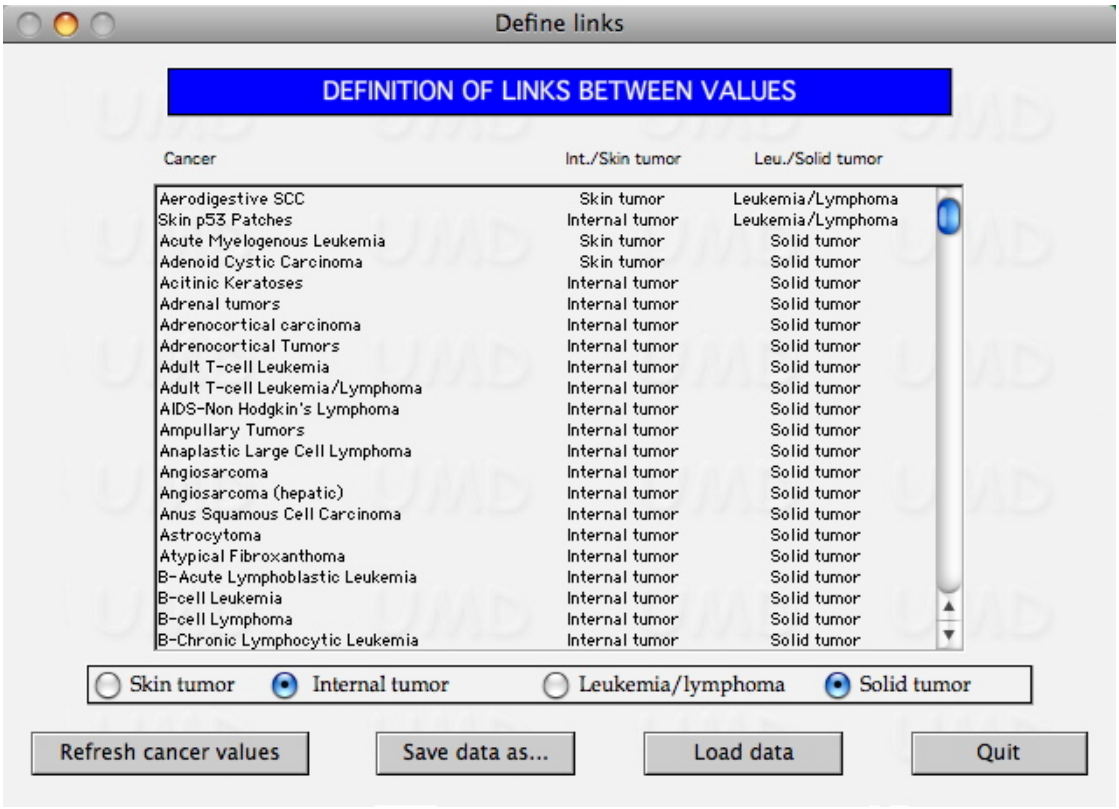
Note that the SIFT predictions are highly dependent of the alignment and a special care should be made to the selection of the various organisms used. The default option is to use the NCBI GI number but this usually gives less efficient predictions.
When the proper SIFT file has been saved, a click on the import button allow the automatic



creation of the corresponding records in the UMD database.
These data can be modified either directly or through another import. They are used by the UMD-Predictor tool to predict the pathogenicity of missense mutations and substitutions.

29) Define links

This function is only useful for genes involved in cancers. It allows to define if cancer entities



are internal or skin tumors and leukemia/lymphoma or solid tumors.

30) Create a new database

This function is used to create a new UMD database (cf. II).

31) Gene type

Allows to define if the gene is involved in cancers or not. If the gene is involved in cancers, this will activate the "*Define links*" function as well as a specific display during the mutation creation.

32) Update clinical data

This function allows to refresh the list of clinical symptoms and associated severities. It is necessary to activate this function prior to the use of the "*Export data*" function.

33) Clinical data type

This function is used to defined the various severities associated to each symptom (text, numeric values, date) (cf.V).

34) Activate clinical search

This function is used to turn on or off the web search module that includes clinical data. It is only useful for on-line UMD-LSDBs.

35) Numeric value ranges

This function is used to defined numerical ranges associated to severities (cf.V).

36) Search preference

This function is used to select the search interface (cf.VII).

37) Splice sites calculation

The user can switch between 2 methods to calculate the acceptor and donor splice sites strength: the UMD algorithm or the Cartegni algorithm. This option can be modified anytime in order to compare predictions.

38) Pathway for dnamlk application

Allows the user to define the directory containing the dnamlk application ued for phylogeny.

39) Activate webserver

This function is used by the Montpellier's team to activate UMD databases through the Internet. Note that we offer free hosting of your UMD database. For more information, please contact Christophe Béroud.

40) Authorize Web Services

This function is used to activate or inactivate Web Services.

41) Activate specific options

This function is restricted to curators for whom specific options have been activated. It requires specific passwords.

42) Specific Web Parameters

This function is restricted to the UMD website administrator.

43) Restricted access

This function is restricted to collaborators for whom specific options have been activated. It requires specific passwords.

44) Web license

This function is used by the Montpellier's team to activate UMD databases through the Internet. Note that we offer free hosting of your UMD database. For more information, please contact Christophe Béroud.

45) 4D_molecules folder

This function is restricted to the UMD website administrator.

46) Password

This function is used to create users and associated passwords. They are used to limit the access to the UMD® software but also through the Internet.

47) Import Passwords

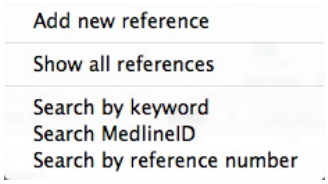
The login and passwords are stored in the UMD® software itself and therefore are lost when the UMD® software is updated. The "Import Passwords" function allows an easy update of passwords from a text file.

48) Quit

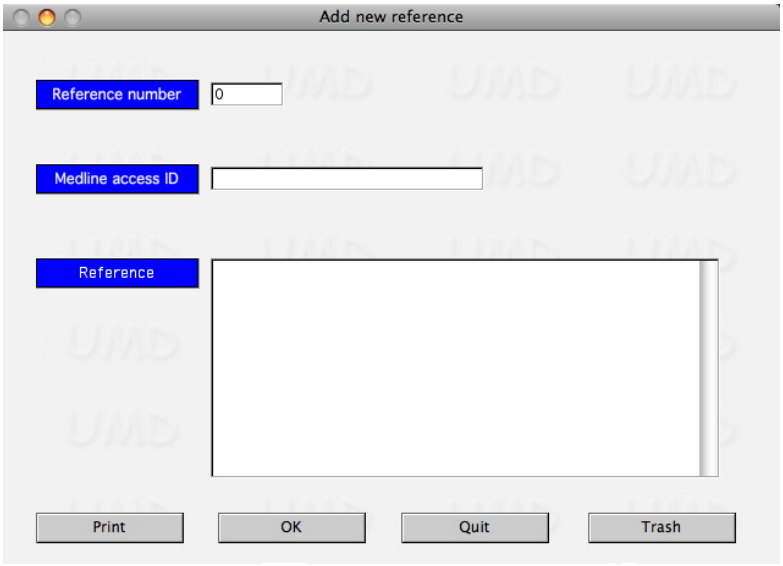
Allows to quit the UMD® software.

VIII-b- The "References" menu

This menu contains 5 functions, which are related to references. In fact, we recommend to link every mutation to a reference in order to facilitate the curation process.



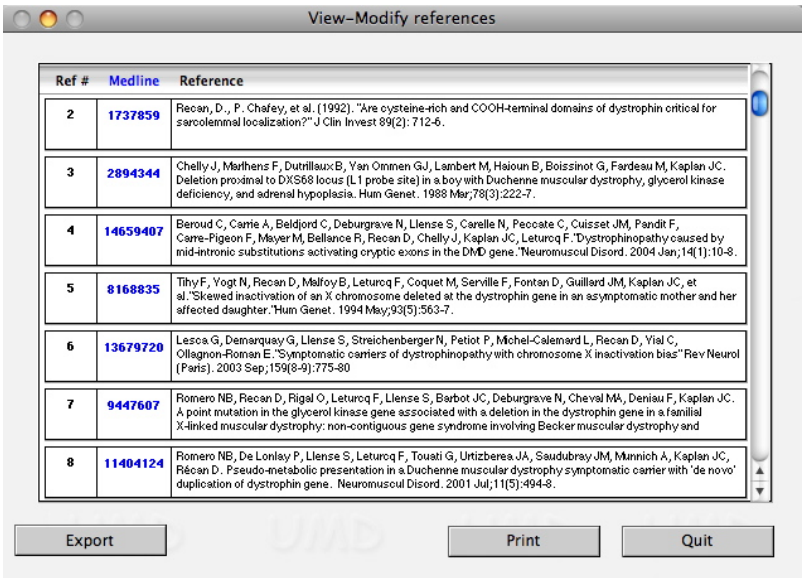
1) Add new reference



This function is used to create a reference. Each reference is characterized by a unique reference number, a Medline access ID (used to establish links with the Medline website) and a reference (free text).

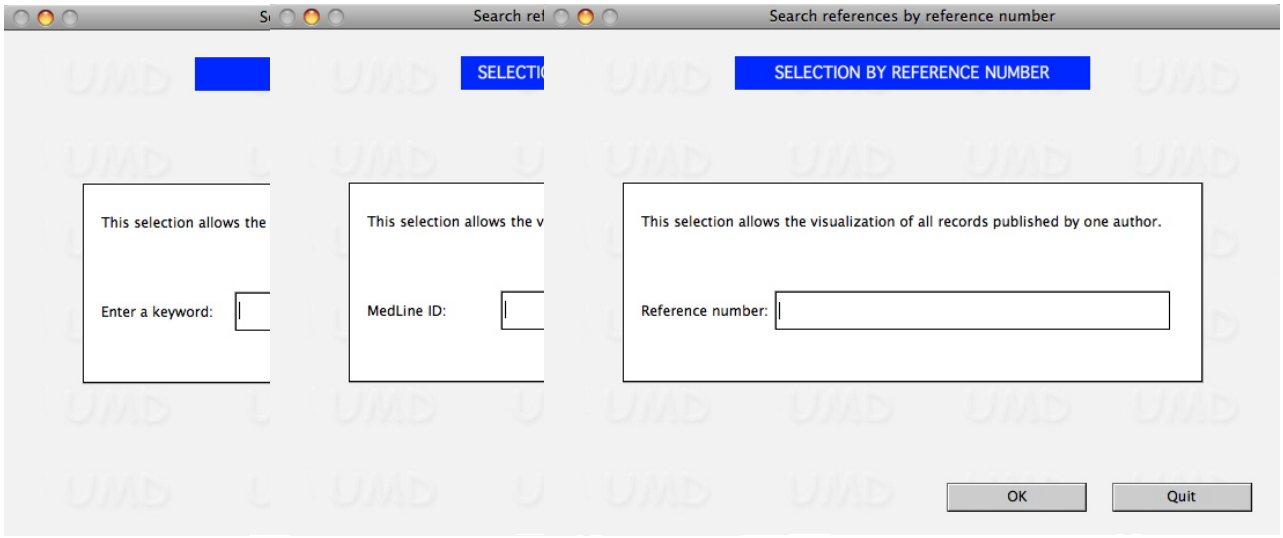
2) Show all references

This function gives access to the list of all references.



3) Search by keyword, Search Medline ID, Search by reference number

These three functions allow to search the reference table using the corresponding criteria (keyword, Medline ID or reference number).

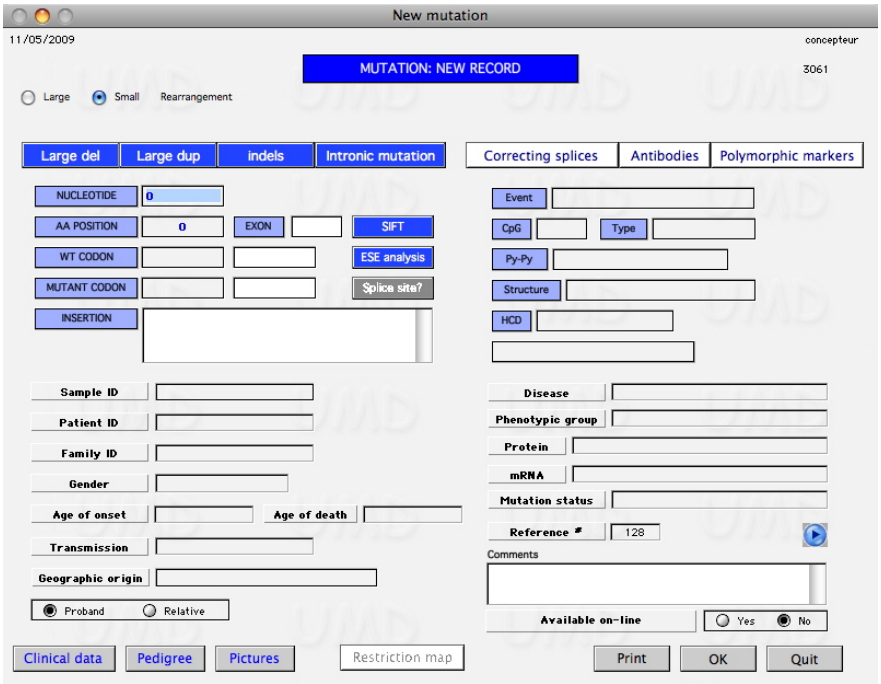


VIII-c- The "Mutations" menu

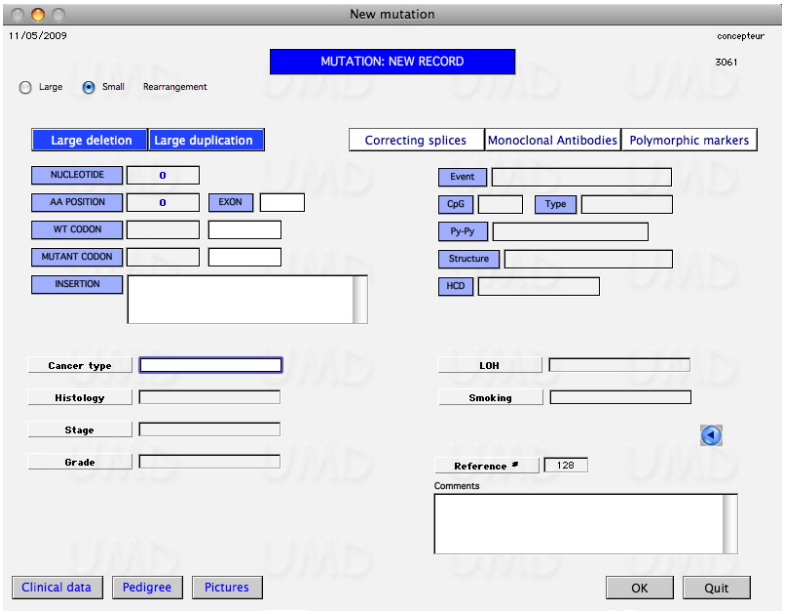
This menu contains 5 functions, which are related to mutations.

1) Add a new record

This is the central function of the UMD® software. It allows the creation of a new mutation. A generic format is used for new databases. Specific input formats could also be used (contact us to create a specific format).

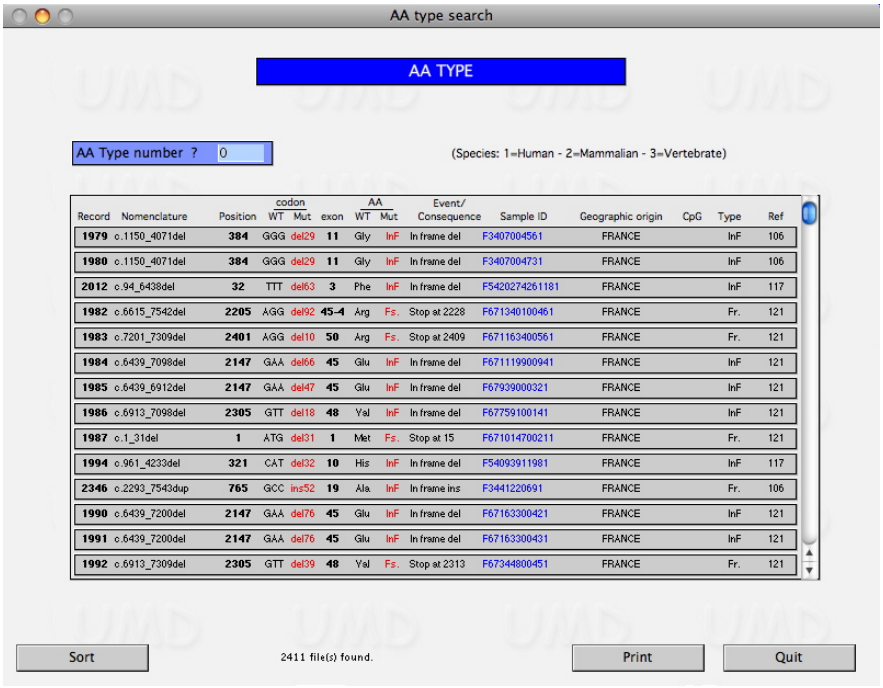


The input format contains fields related to the mutation itself (top part, cf. III), to the patient/sample (bottom part). The Blue arrow gives access to the cancer related input format:



2) AA type search

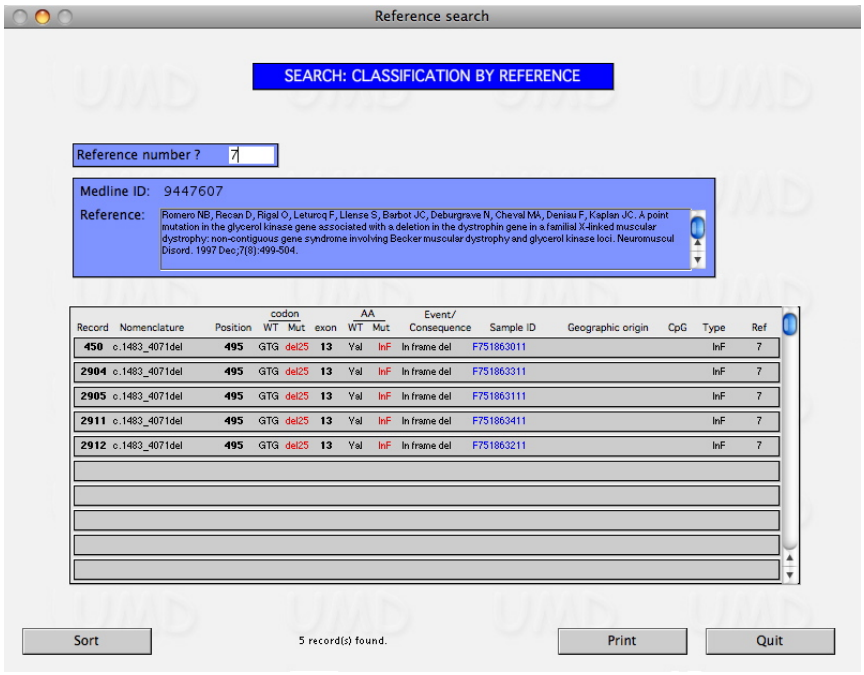
Each amino acid is associated with a "Species" field. This data could be used to search for mutations involving amino acids conserved among vertebrates, mammals or specifically



human. This function uses the generic display format.

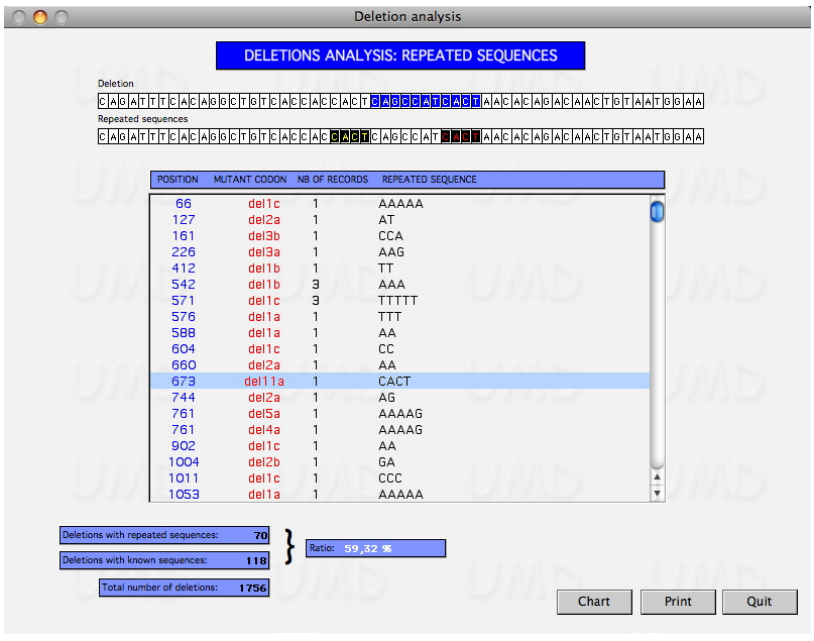
3) Reference search

This function allows a rapid display of mutations associated to a specific reference. It is frequently used for curation in order to avoid duplicates from a specific team.



4) Deletion analysis

Small deletions frequently result from slippage of the polymerase during the replication (for example deletion of an A in a stretch of As) or to a more complex process that involves repeated sequence and lead to the deletion of the sequence localized between the repeats with or without the repeats themselves.
During a mutation's creation, the UMD® automatically searches for repeated sequences that are localized at the extremities of the deletion and for stretches of nucleotides that could explain the molecular mechanism that led to the mutation. These data are accessible via the "*Deletion analysis*":

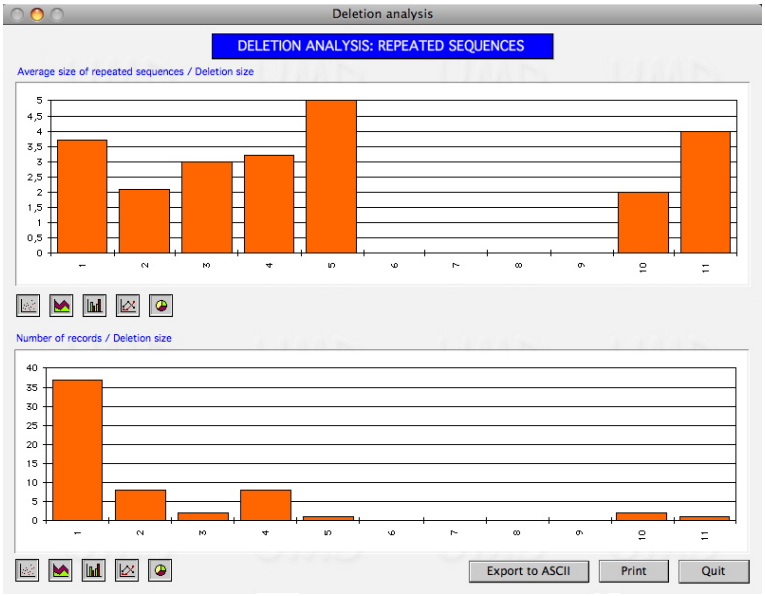


In the upper part of the screen are displayed the reference and the mutant sequences. The deleted sequence is shown in white on a blue background while repeated sequences are

displayed in yellow and red on a black background. A simple click on the table will display the corresponding information on the upper part. Simple statistics are also available (bottom). A "*Chart*" button gives access to two graphics:

- The first one displays the average size of repeated sequences compared to the deletion size;
- The second one displays the number of records found for each deletion size.

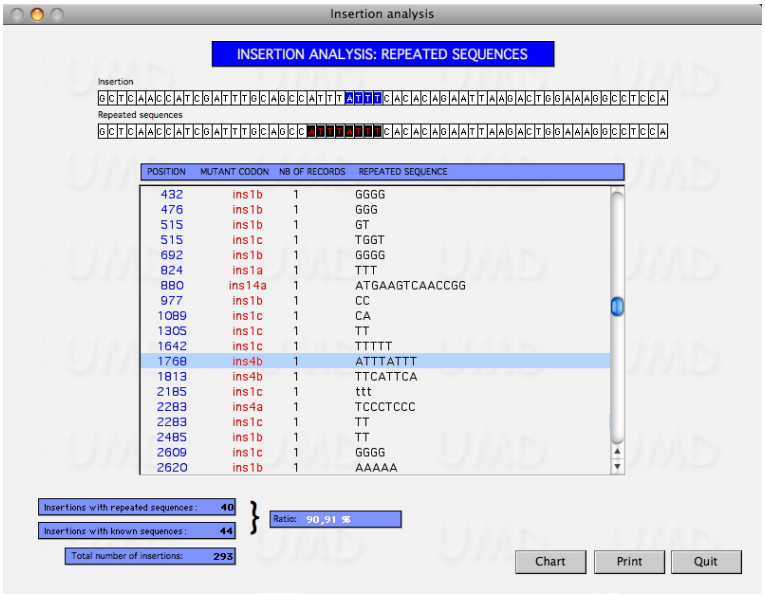
Both graphs can be modified using dedicated buttons, and corresponding data could be exported as an text tab-delimited file in order to be used in more sophisticated graphical



display tools.

5) Insertion analysis

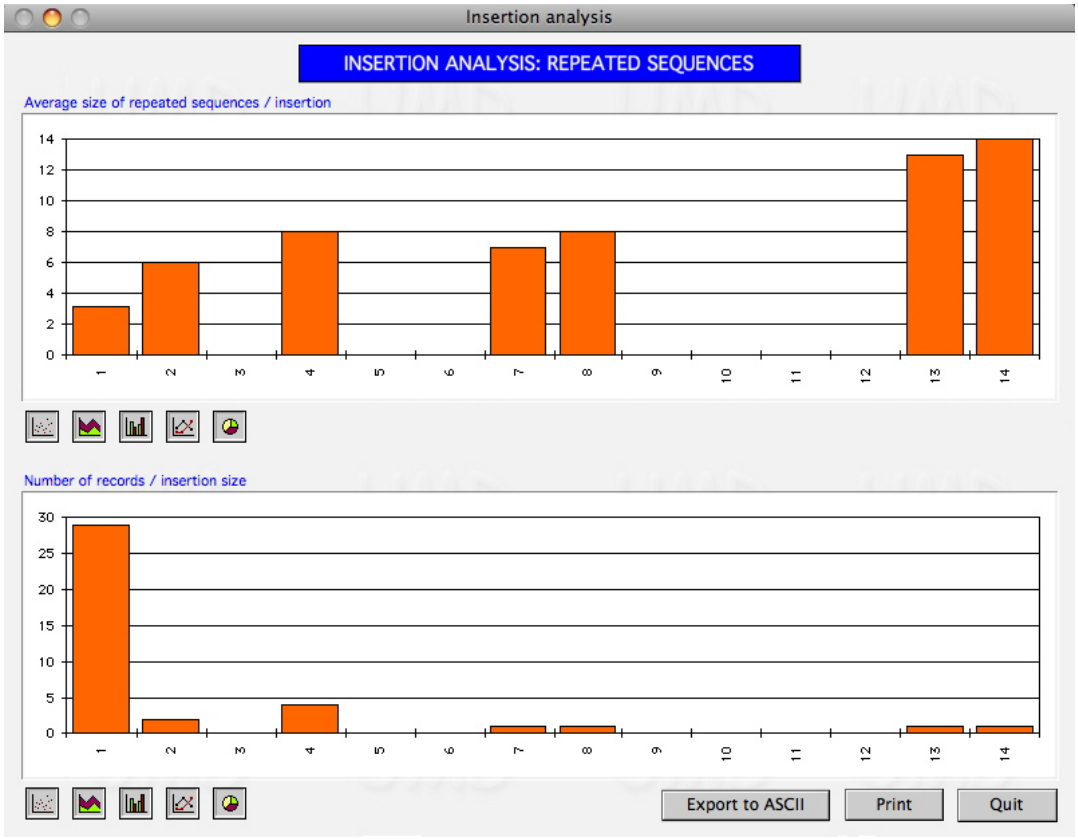
As for small deletions, small insertions frequently result from the creation of repeated sequences. Therefore during a mutation's creation, the UMD® automatically searches for repeated sequences generated by the mutation:



In the upper part of the screen are displayed the reference and the mutant sequences. The inserted sequence is shown in white on a blue background while repeated sequences are displayed in red on a black background. A simple click on the table will display the corresponding information on the upper part. Simple statistics are also available (bottom). A "*Chart*" button gives access to two graphics:

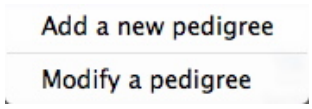
- The first one displays the average size of repeated sequences compared to the deletion size;
- The second one displays the number of records found for each deletion size.

Both graphs can be modified using dedicated buttons, and corresponding data could be exported as an text tab-delimited file in order to be used in more sophisticated graphical display tools.



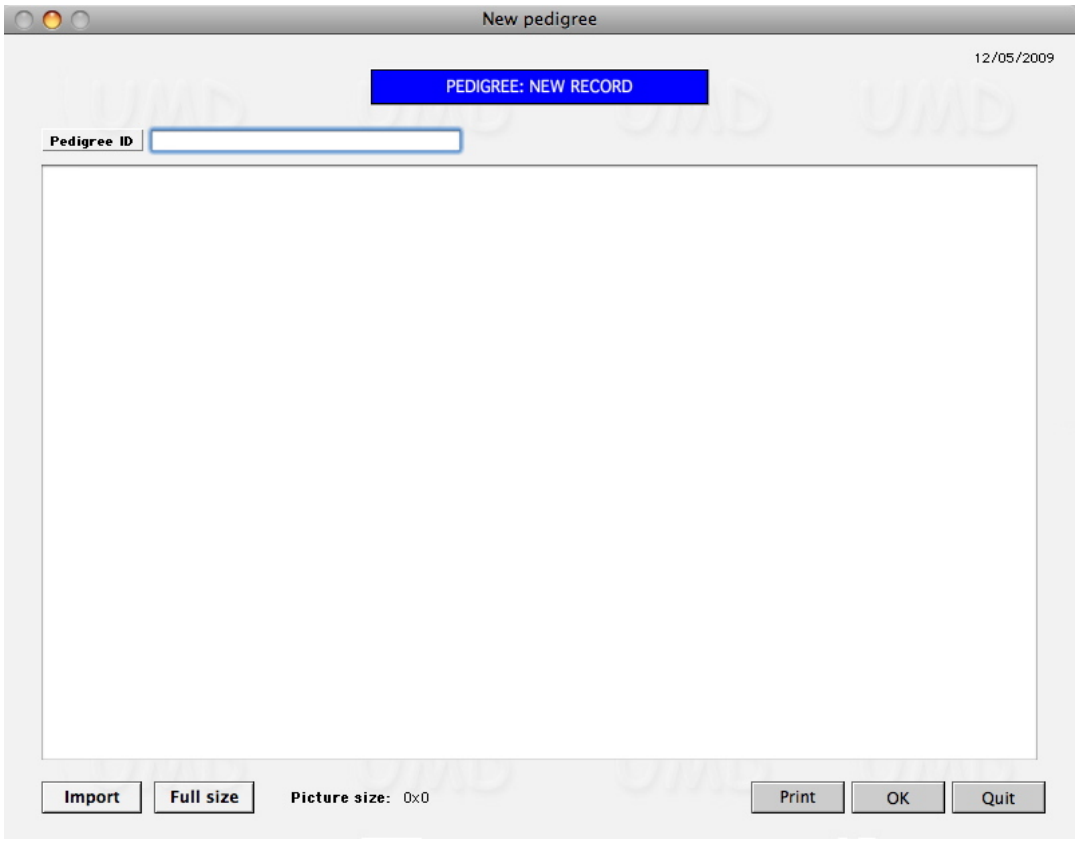
VIII-d- The "Pedigree" menu

This menu contains 2 functions, which are related to the creation and modification of pedigrees.



1) Add a new pedigree

This function allows the import of a pedigree file. Each pedigree should be labelled with a unique identifier (pedigree ID). If the pedigree picture is larger than the display format, a click on the "Full size" button will reformat the display screen in order to view the picture in



the most appropriate format.

2) Modify a pedigree

This function is used to list all pedigrees and modify a specific record (update or delete).

VIII-e- The "Show all mut." menu

This menu contains 4 functions, which are related to the display of mutations and queries.

1) View-Modify mutations

This function gives access to the list of all mutations recorded in the database (cf. VI-b).

2) Define a display format

Multiple personalized display formats can be created in the UMD® software (cf. VI-a).

3) View-Modify display formats

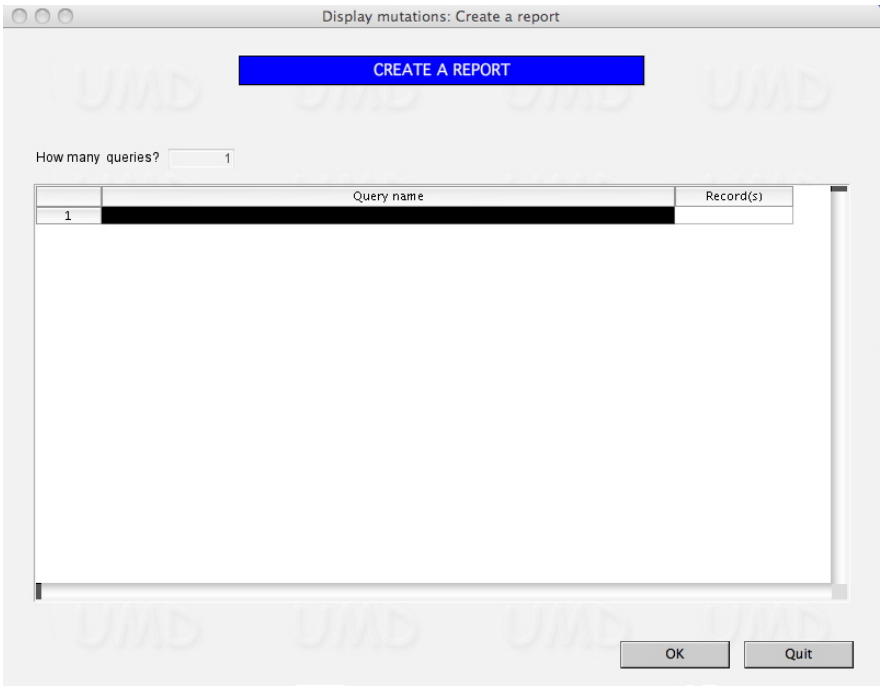
This function gives access to the modification of personalized display formats (cf. VI-b).

4) Free search

With this function, the user can select mutations records using one or many criteria. He can use two search interfaces to perform the queries (cf. VII).

5) Create a report

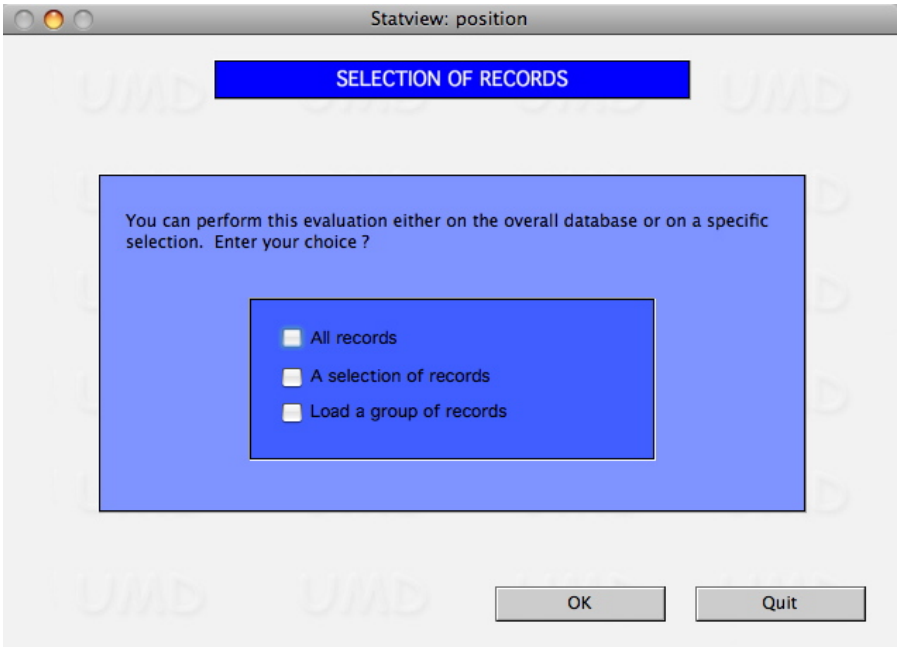
This function is used to compare the number of records corresponding to various selection criteria. The user can define the number of lanes and enter a title for each selection criteria. A click on the lane number gives access to the search interface and returns the number of records.

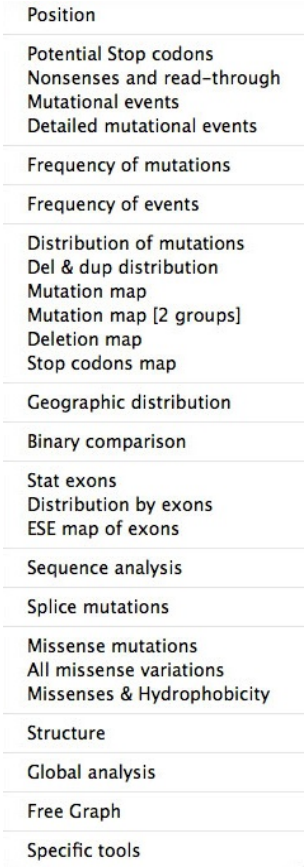


VIII-f- The "Statview" menu

This menu contains 27 functions, which are related to the data analysis either using statistical or graphical presentations.

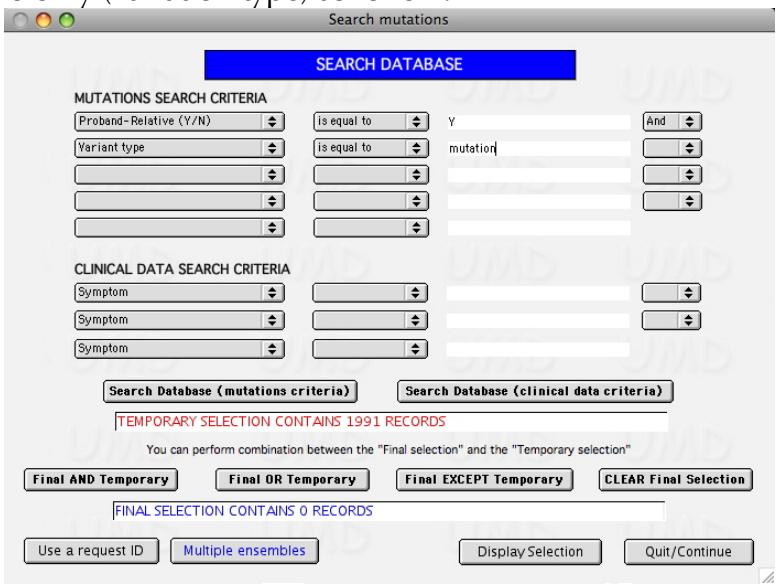
Because it is very useful to be able to perform analysis on a subset of records, all functions from this menu are linked to a "Selection of record" system.





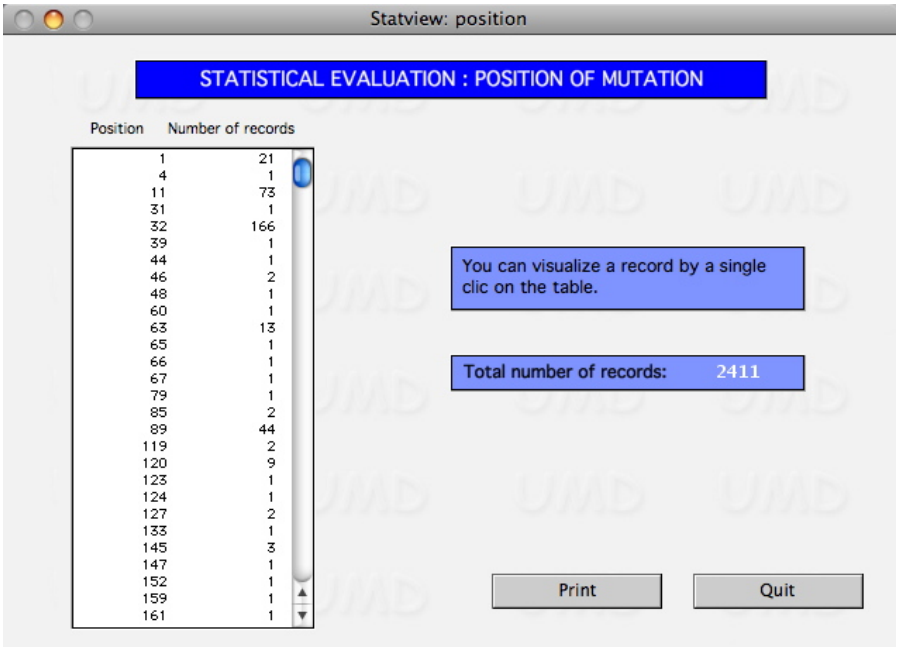
The user can work from "*All records*", freely select a group of records using the search interface when selecting the "*A selection of records*" option or use a previously saved ensemble of records using the "*Load a group of records*".

Note that working with all records should be avoided if you want to study molecular events. In this situation you need to select probands using the search interface. You can also wish to filtrate for mutations only (variation type) as follow:



1) Position

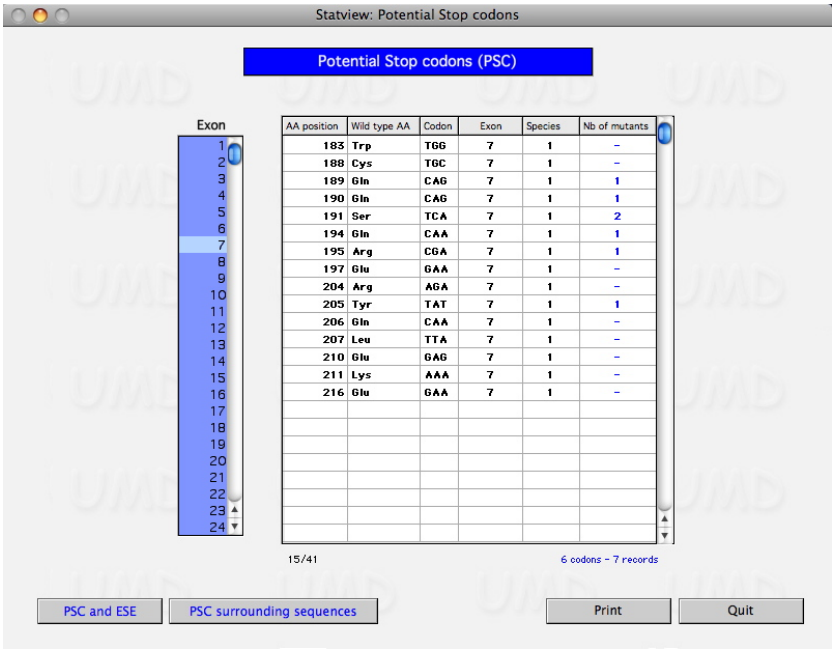
This function gives access to the number of records (mutations) localized at a specific amino acid residue. If large rearrangements are involved, they are associated to the first residue targeted by the mutation.



A simple click on one element of the list will open a window with corresponding records.

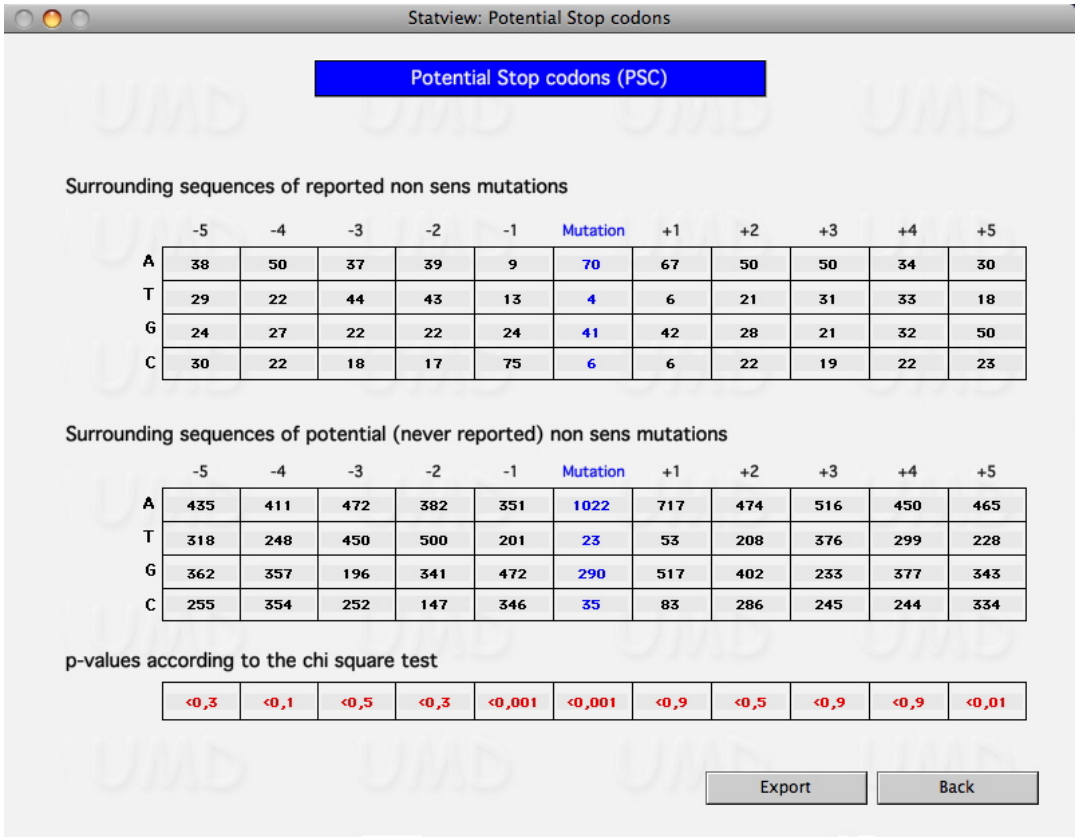
2) Potential stop codons

This function gives for each exon, the list of codons that can be mutated in premature stop codons (PTC) by a single substitution and the number of such mutations.



In this example, 15 out of 471 codons of the exon 7 of the DMD gene could lead to PTC. Among these 7 PTC, mutations have been reported for 6 of them.
Two buttons lead to complementary data:

The "*PSC surrounding sequences*" option gives statistical data about the environment of observed PTC compared to potential stop codons for which no mutation has never been reported. This option is used to study the environment of PTC and find out if this context could play a role in PTC efficiency via codon recognition and read-through or other mechanisms.



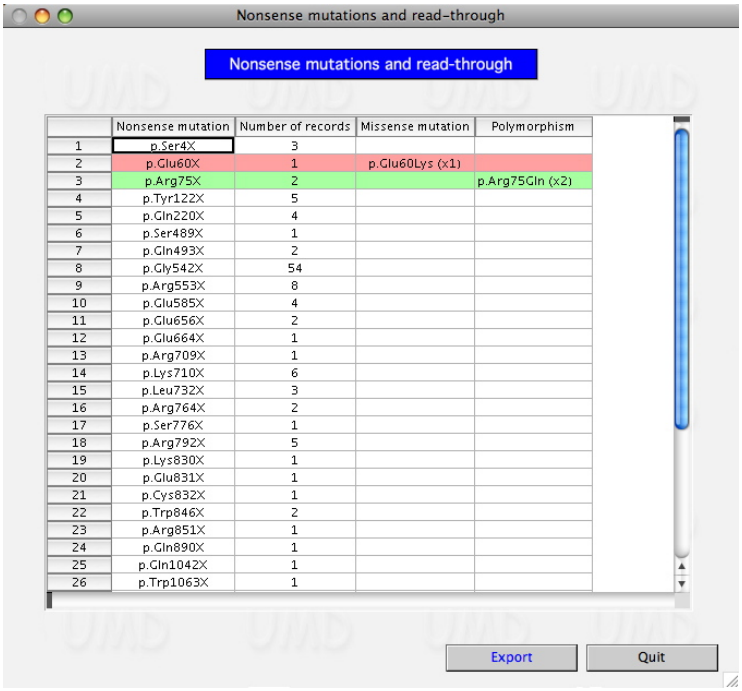
	-5	-4	-3	-2	-1	Mutation	+1	+2	+3	+4	+5
A	38	50	37	39	9	70	67	50	50	34	30
T	29	22	44	43	13	4	6	21	31	33	18
G	24	27	22	22	24	41	42	28	21	32	50
C	30	22	18	17	75	6	6	22	19	22	23

	-5	-4	-3	-2	-1	Mutation	+1	+2	+3	+4	+5
A	435	411	472	382	351	1022	717	474	516	450	465
T	318	248	450	500	201	23	53	208	376	299	228
G	362	357	196	341	472	290	517	402	233	377	343
C	255	354	252	147	346	35	83	286	245	244	334

<0,3	<0,1	<0,5	<0,3	<0,001	<0,001	<0,9	<0,5	<0,9	<0,9	<0,01

3) Nonsenses and read-through

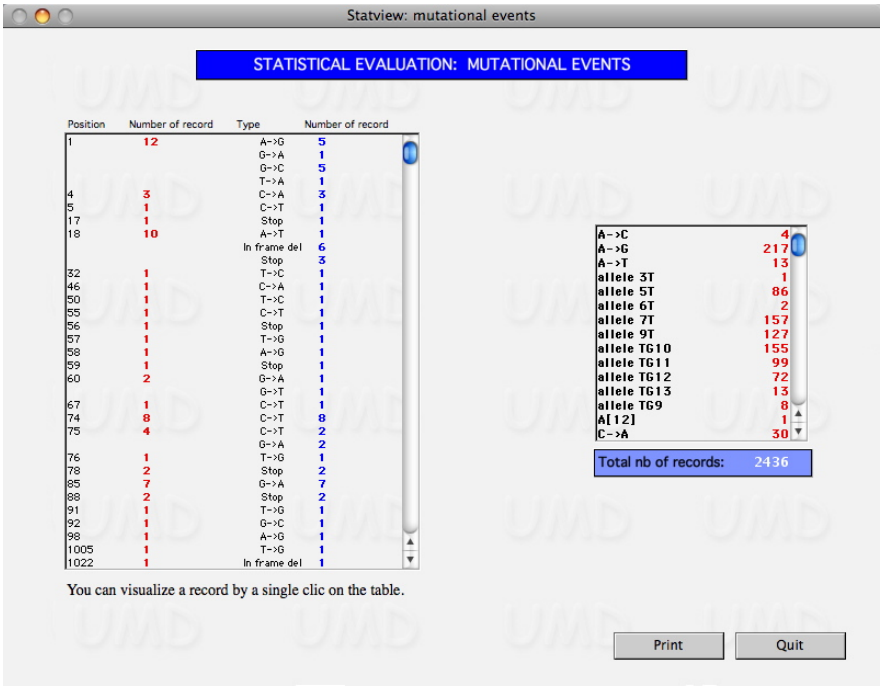
This function is used to evaluate the potential to correct a premature termination codon mutation by a read-through therapeutic strategy. In fact, read-through can be induced by various chemical compounds (aminoglycosides, PTC124) and leads to the incorporation of an amino acid at the premature stop codon position. This will therefore lead to the production of a quasi normal protein with an amino acid substitution at the position of the premature stop codon.
This function therefore lists all reported nonsense mutations with for each position, the number of pathogenous missense mutations and/or polymorphisms (non-pathogenic mutations) affecting the residue.
Only nonsense mutations affecting a residue without associated pathogenic missense mutations are good candidates for nonsense read-through strategies. Among these nonsense mutations, those associated to polymorphisms are even better candidates for these therapeutic strategies.



In this example (CFTR gene), the p.Glu60X targets the Glu 60 residue that is also targeted by a missense mutation and therefore not a good candidate. Conversely, the p.Arg75Gln polymorphism targets the Arg 75 residue involved in the p.Arg75X and is therefore a better candidate for nonsense read-through.

4) Mutational events

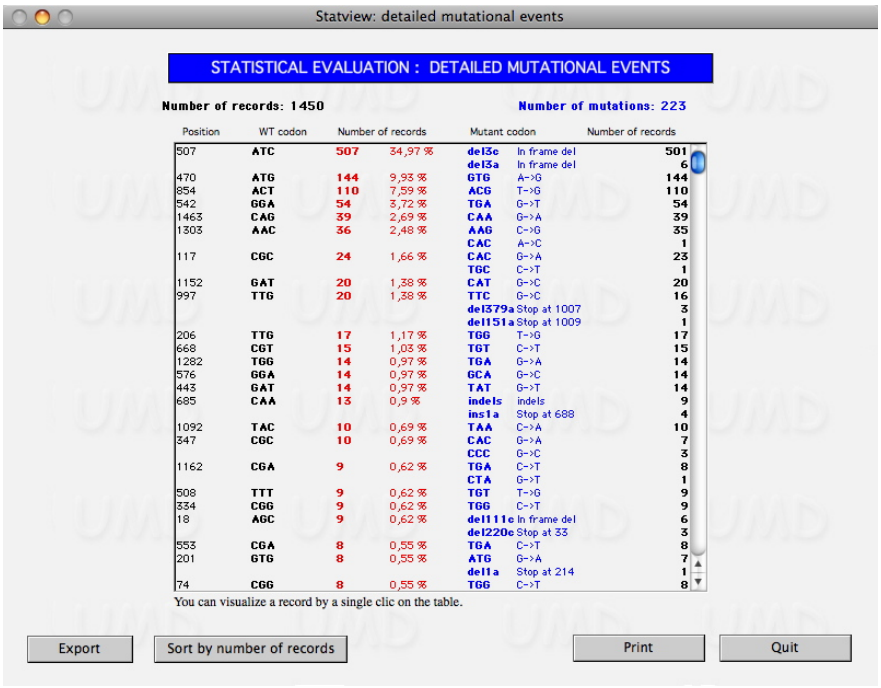
This function list mutational events reported at each codon. A summary table is also available. A click on one row of the tables allows the display of corresponding records using



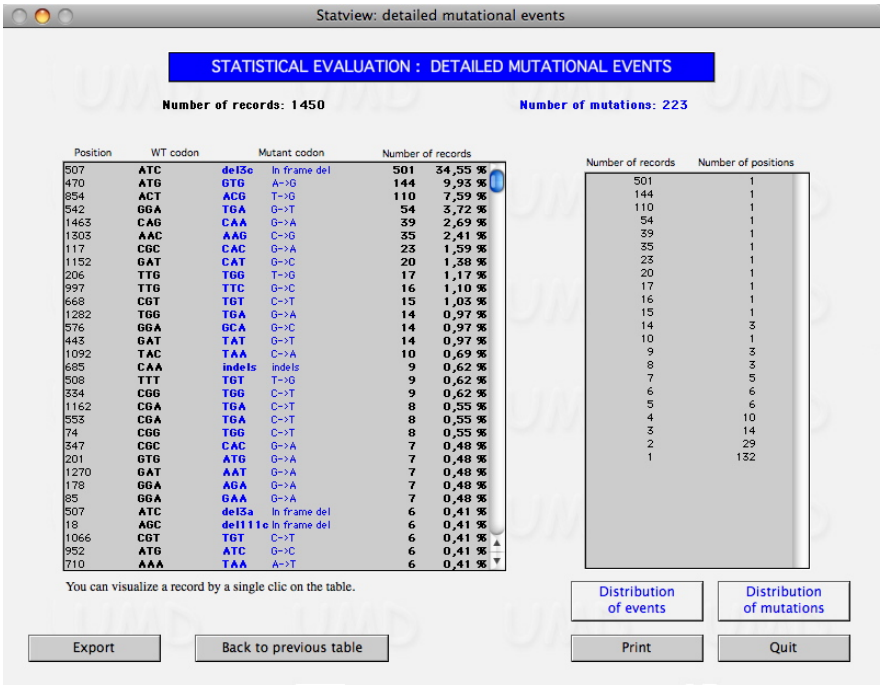
the personalized default display format.

5) Detailed mutational events

This function displays all informations about mutational events. Positions are sorted by frequency.

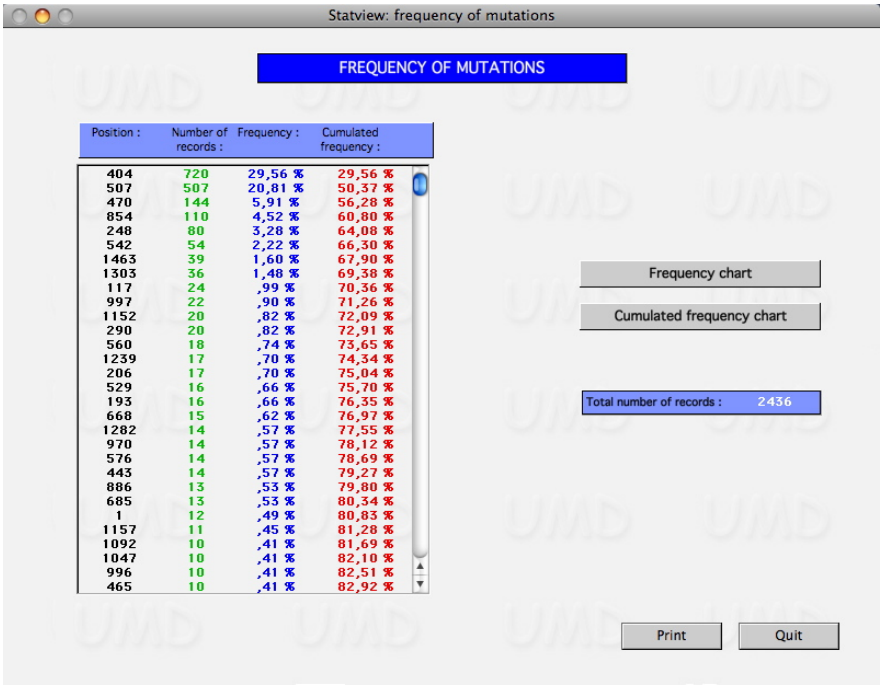


A click on the "*Sort by number of records*" button opens a second window that sorts mutational events by frequency. A second table also lists the number of positions with similar number of records.



6) Frequency of mutations

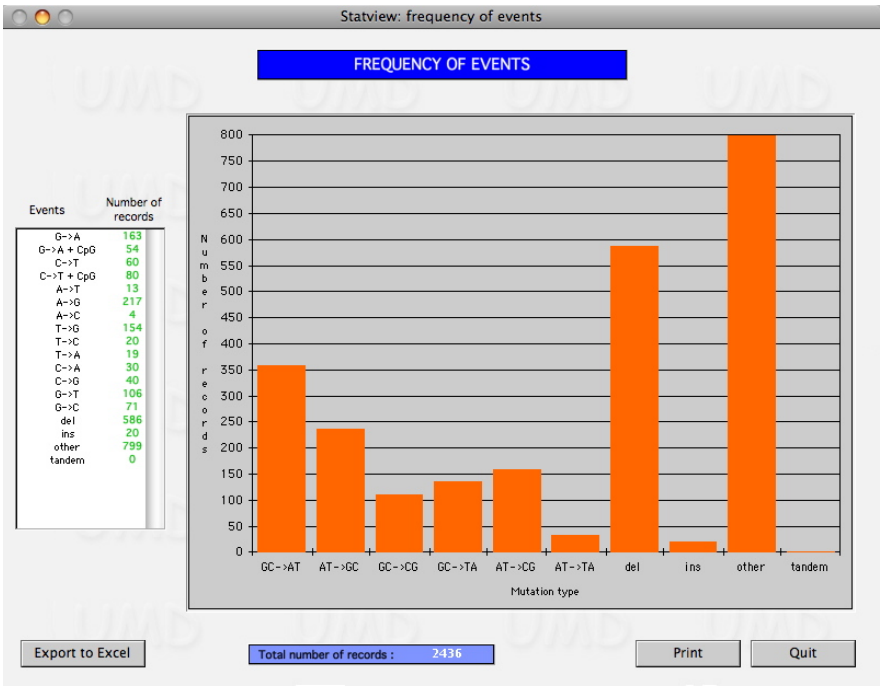
This function sorts the amino acid positions according to the number of records. In addition, a cumulative frequency value allows the evaluation of screening strategies and their relative efficiency. This function is frequently used on subsets of records selected by phenotypes.



Two graphical displays are also available to display positions by frequency of mutations or by cumulated frequency of mutations.

7) Frequency of events

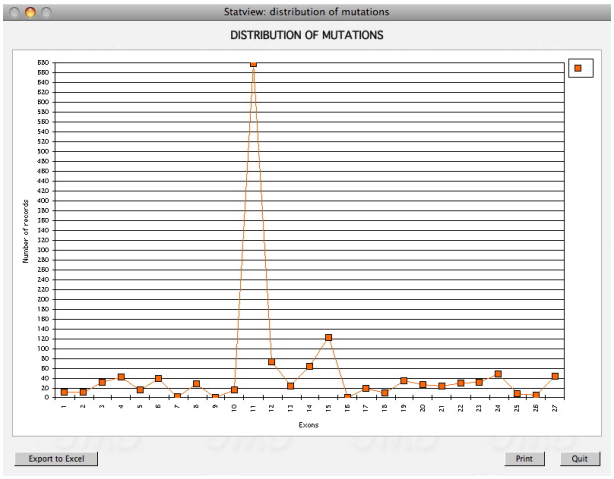
This function displays the distribution of mutational events.

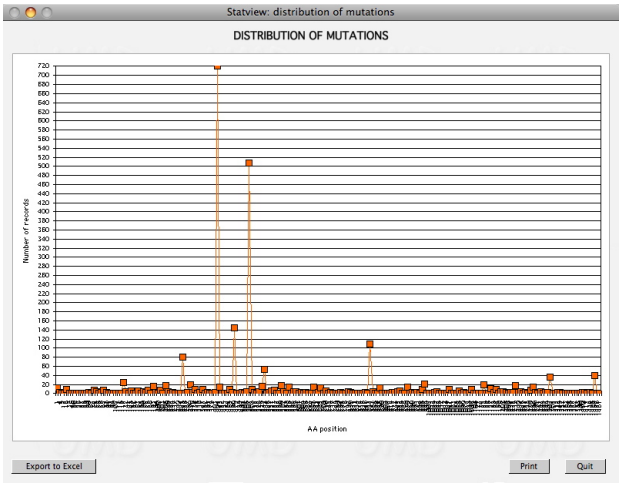


8) Distribution of mutations

This function is designed to compare the mutations of up to 8 groups of records. Each group is selected using the regular search interface. The results are displayed in 2 formats:

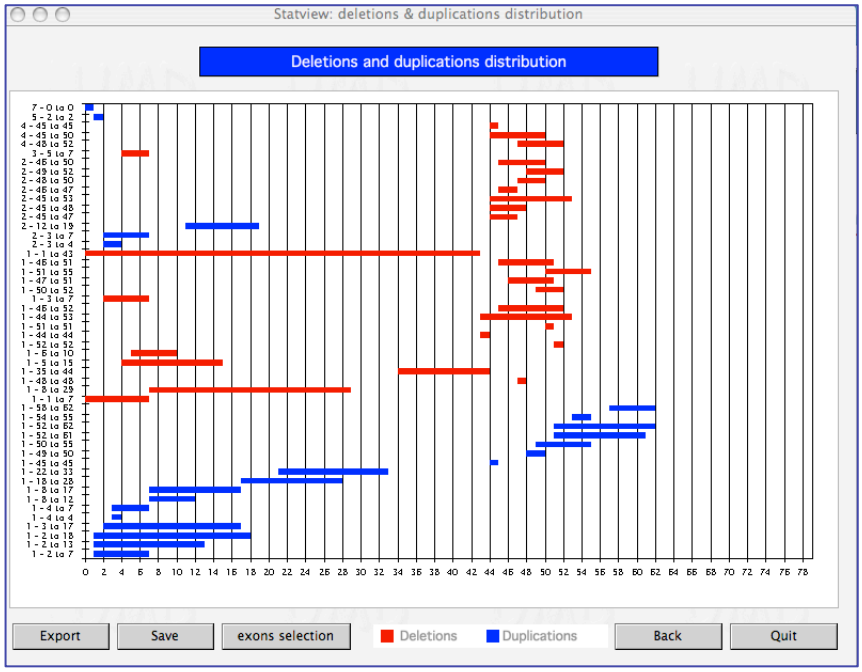
- A graphical display of the distribution of mutation per exon;
- A graphical display of mutations by residue.

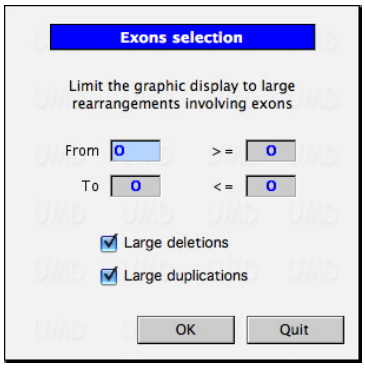




9) Del & dup distribution

This function is restricted to the study of large deletions and duplications (≥ one exon). The user can select the exons of interest prior to the display.





In the x-axis are displayed the various exons and in the y-axis the various deletions/duplications with the corresponding number of records. The "Exon selection" button allows the modification of display parameters.

10) Mutation map

After choosing the subset of records that should be used for this analysis, the user can choose between two display options:

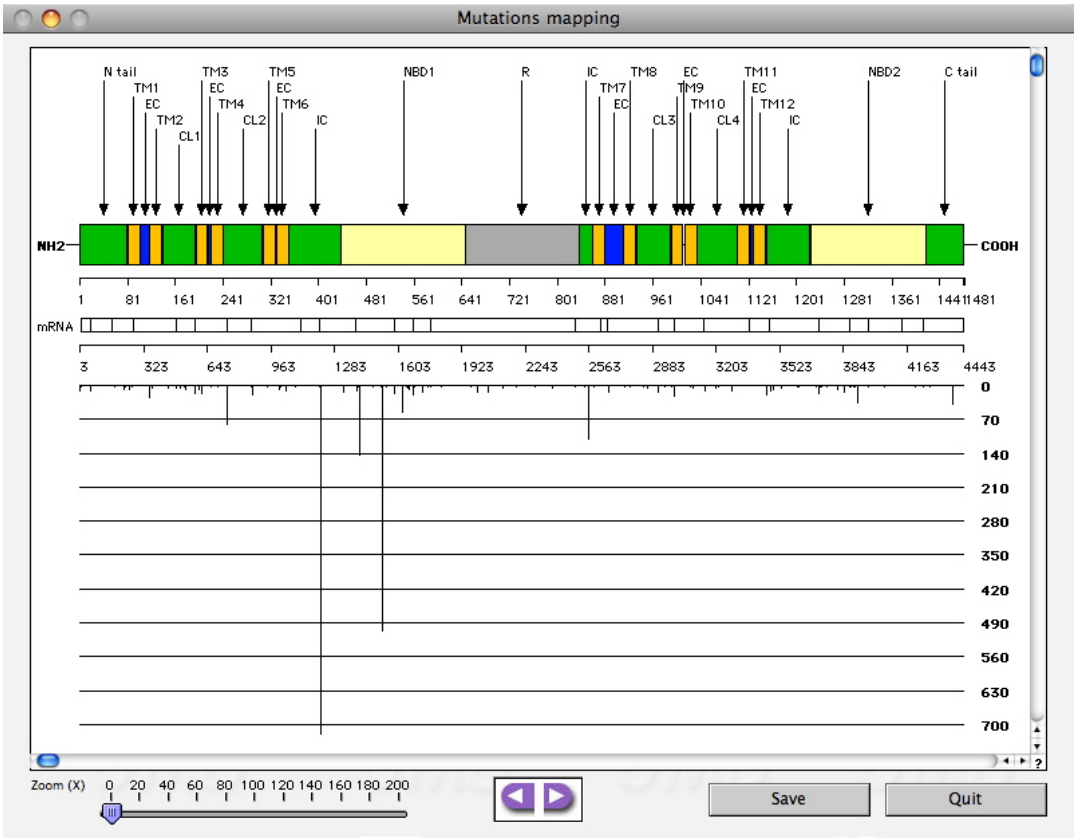
"All" is used to display all mutations and their relative numbers

"Phenotypes" is used to display residues harboring mutations for each phenotype. In order to avoid differences between phenotypes frequency, the number of mutations is not presented in this graph.

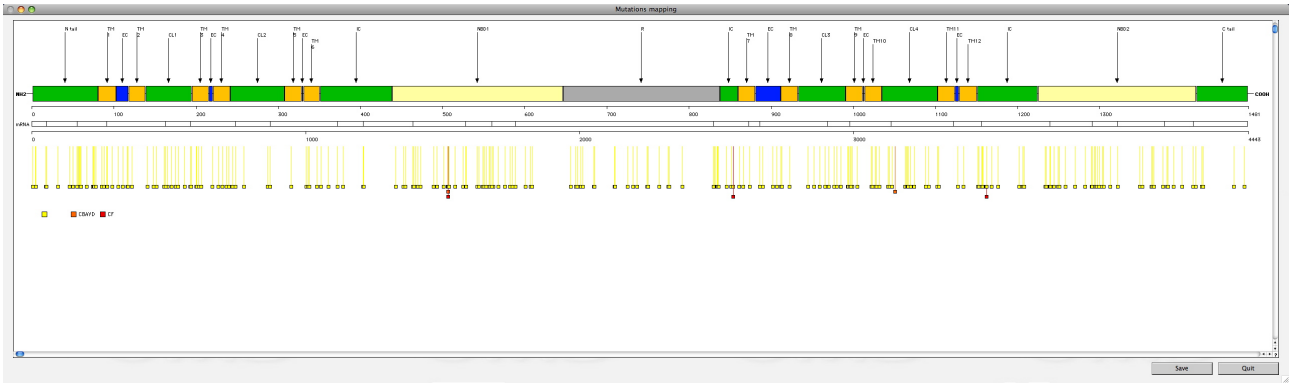
Both graphs use the same scales: Protein structural domains and associated amino acid scale (top) and the various eons and the associated nucleotide scale (middle).

Note the various structural domains are labelled with short names and used colors defined with the "Structure preference" function from the "File" menu (cf.).

Below are presented either the number of mutations at each residue or the presence or absence of a mutation at each residue for each phenotype.

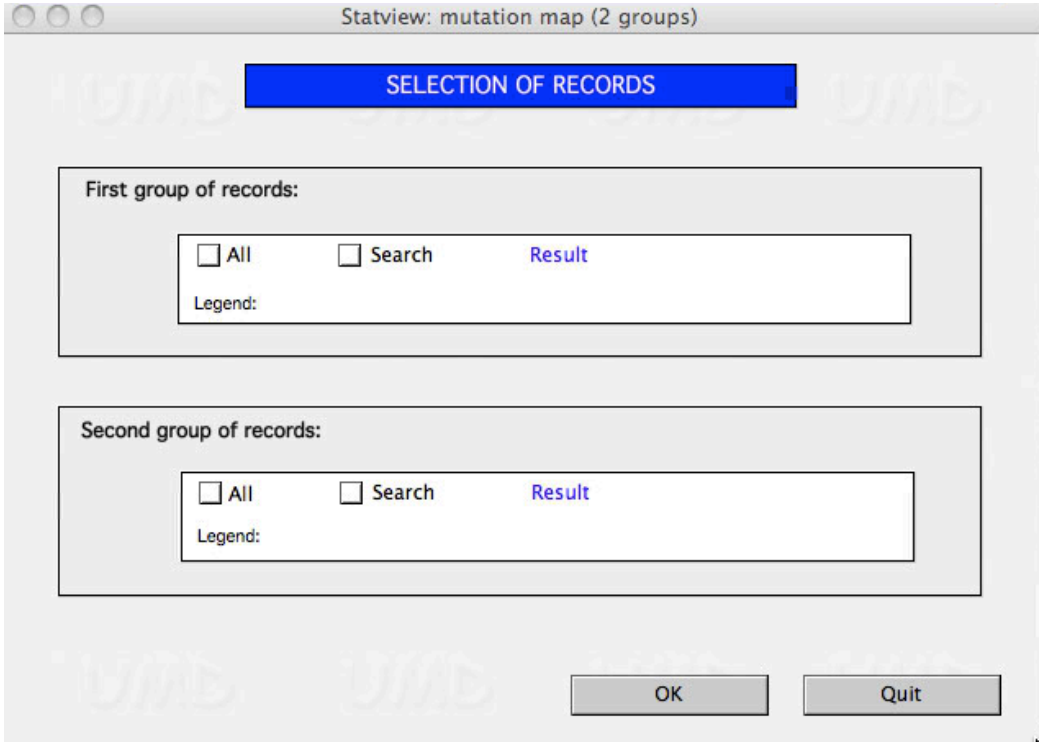


Both pictures can be saved as pictures using the "*Save*" buttons.
In the "All" display format, zooming options are available (up to x200). The user can then click on the graph to select the region of interest or can navigate within the graph using the two violet arrows.



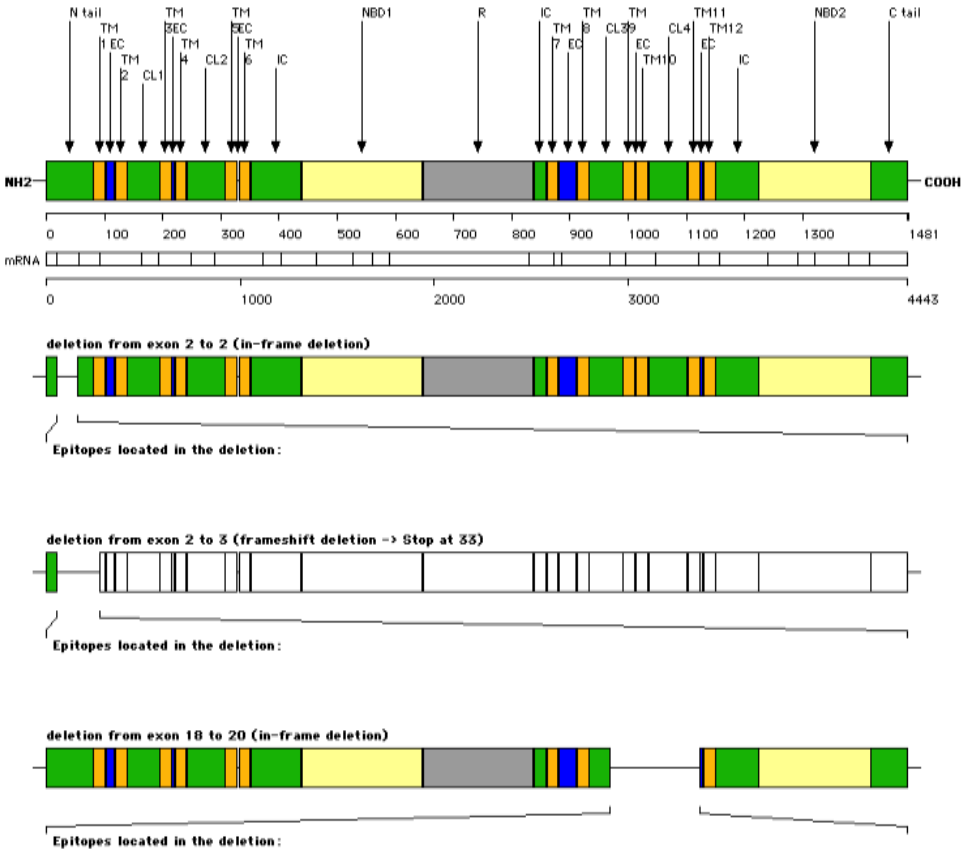
11) Mutation map [2 groups]

This function is similar to the "Mutation map" except that it allows the user to compare two subsets of records. Unlike the other function, the user cannot choose the display options but he can add a legend for both subsets of records.



12) Deletion map

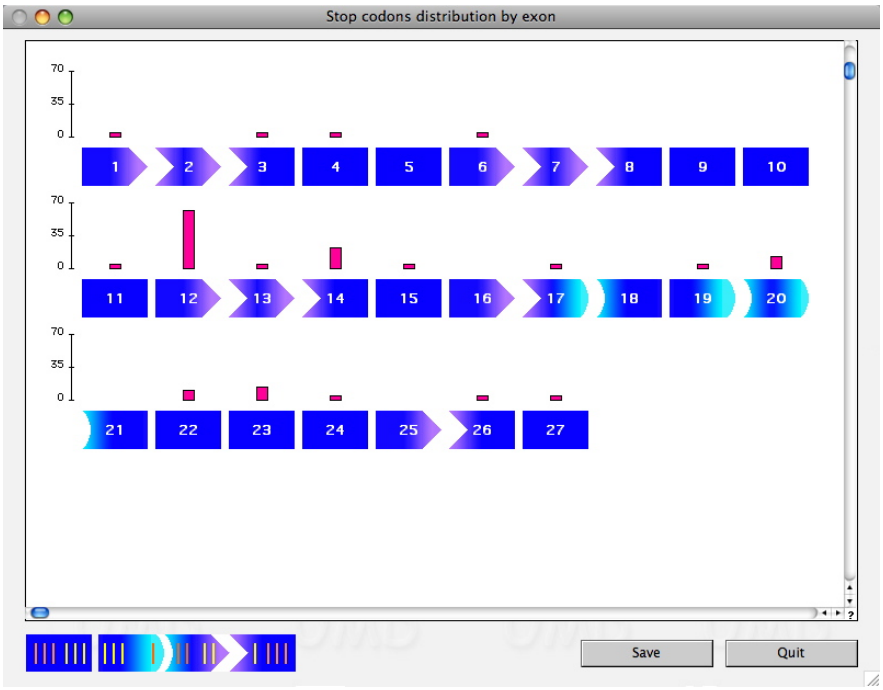
This function displays the consequence of large deletions on the reading frame and the various antibodies epitopes localized within the deletion.



In this example, the deletions of exon 2 and exons 18 to 20 of the CFTR gene are in-frame deletions and therefore lead to internal deletions. The various structural domains affected by these mutations could easily be determined with the color code system.
The deletion of exons 2 and 3 is out-of-frame and probably lead to the absence of protein. Nevertheless, the predicted potential truncated protein is displayed in the graph.

13) Stop codons map

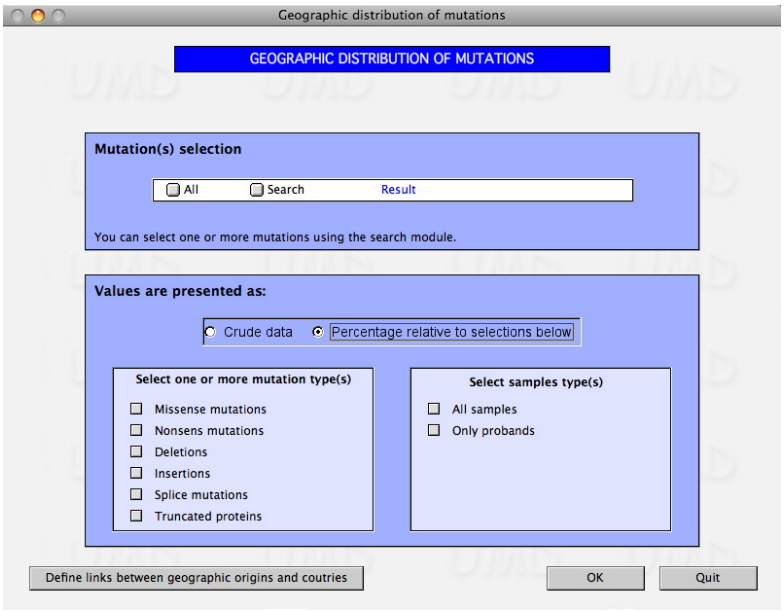
This function displays the exon phasing and the position and number of reported nonsense mutation (stop codons). this function was primarily designed to choose the most relevant



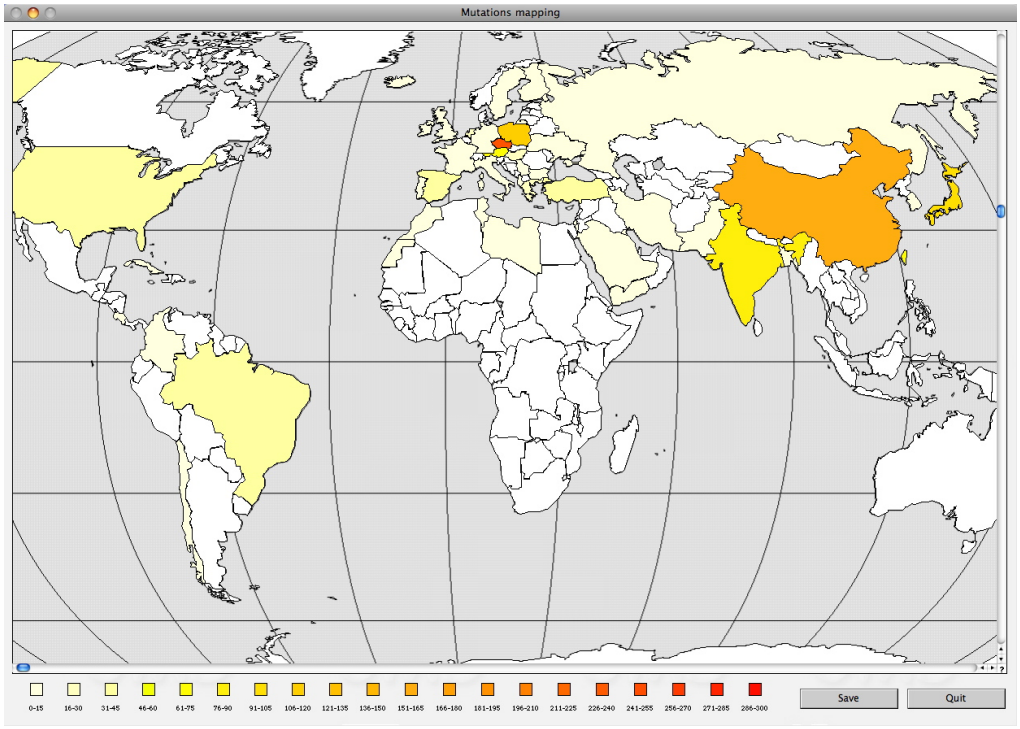
exons that could rescue nonsense mutations by the exon-skipping therapeutic approach.

14) Geographic distribution

This function can only be used if data about geographic origin of patients have been collected. The user can choose all or a subset of records, select one or more mutation types and to display crude data or percentage relative to the selected records.

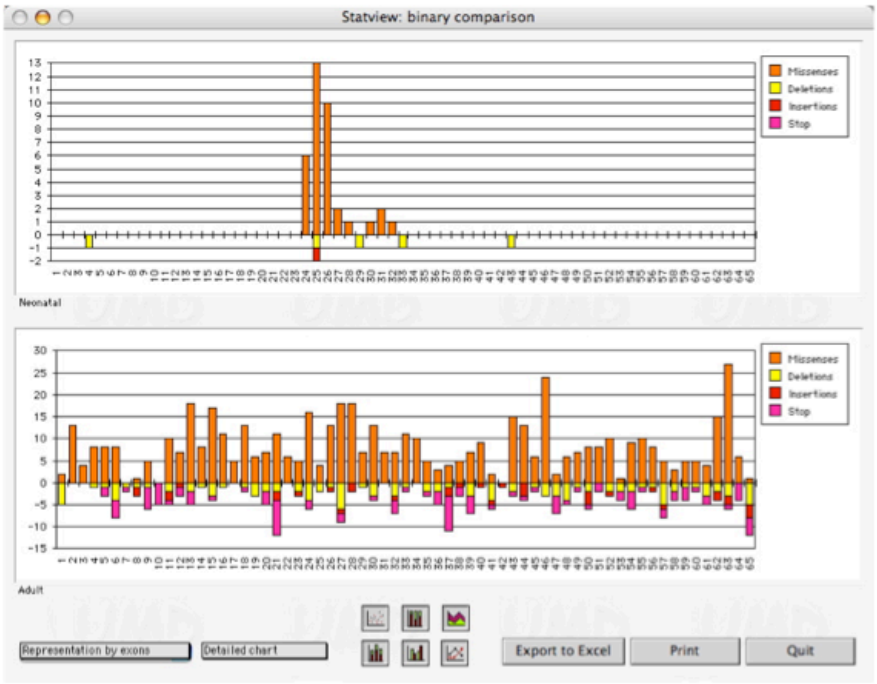


The example below displays the distribution of missense mutations within the ATP7B database.



15) Binary comparison

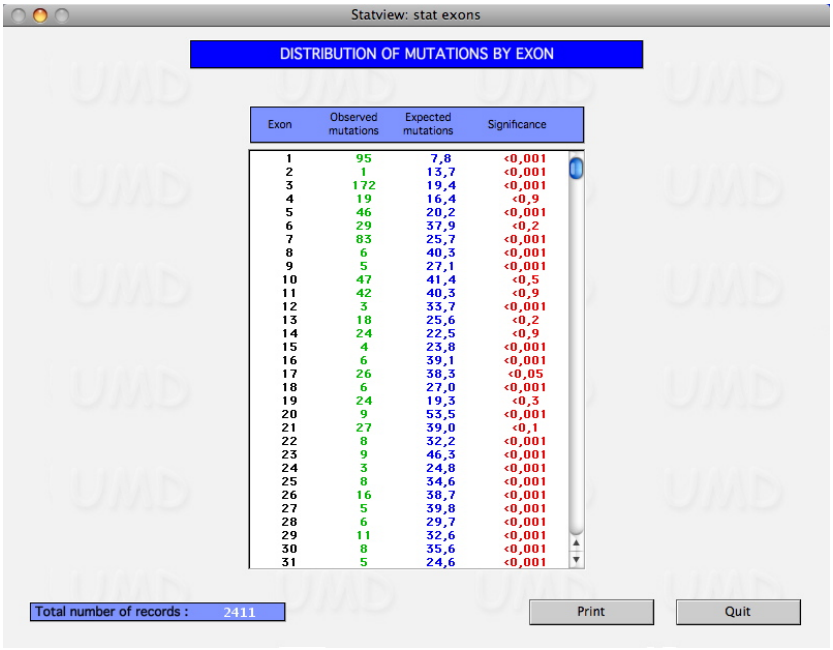
This function is used to compare two sets of data. In this example the neonatal form of the



Marfan syndrome (top) to the Adult form of the disease.
Note that the user can use different parameters: representation by residues or exons; detailed (missense, deletions, insertions and nonsense mutations) or cumulated charts as ell as various display options. Finally, the corresponding data can be exported "*Export to Excel*" button.

16) Stat exons

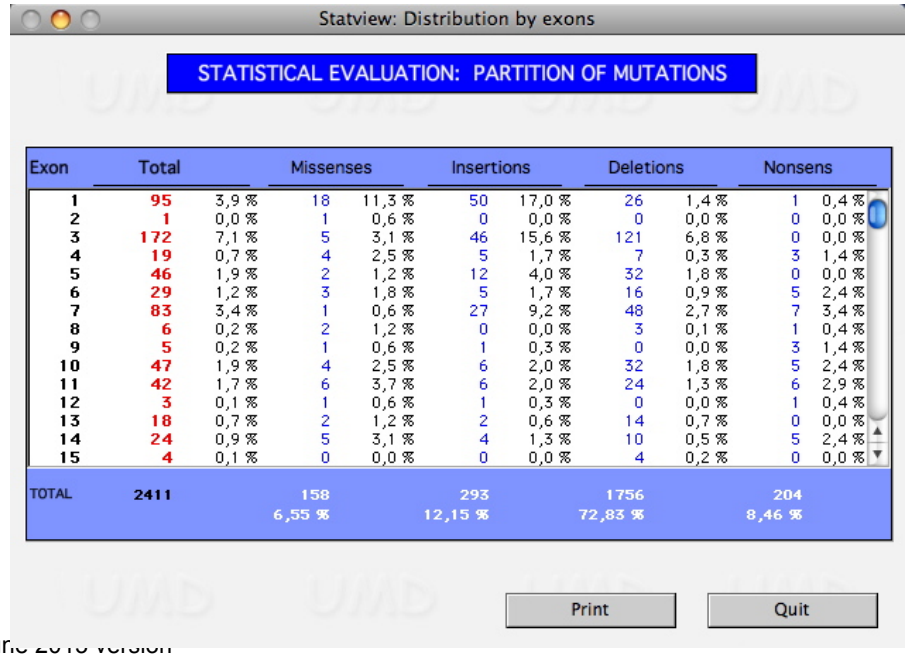
This function is used to look for hotspots associated with specific exons. The UMD® software analyzes each exon to define the expected mutation number corresponding to the hypothesis of random distribution of mutations. Are taken into account the exons lengths and



their nucleotides compositions to take into account the degenerated code.
After selection of records (see introduction of this chapter), the observed and expected number of mutations per exon are displayed as well as the statistical significance (Chi-square test).

17) Distribution by exon

This function is complementary to the previous one. It gives details about mutation types found in each exon (missense, insertions, deletions and nonsense mutations).

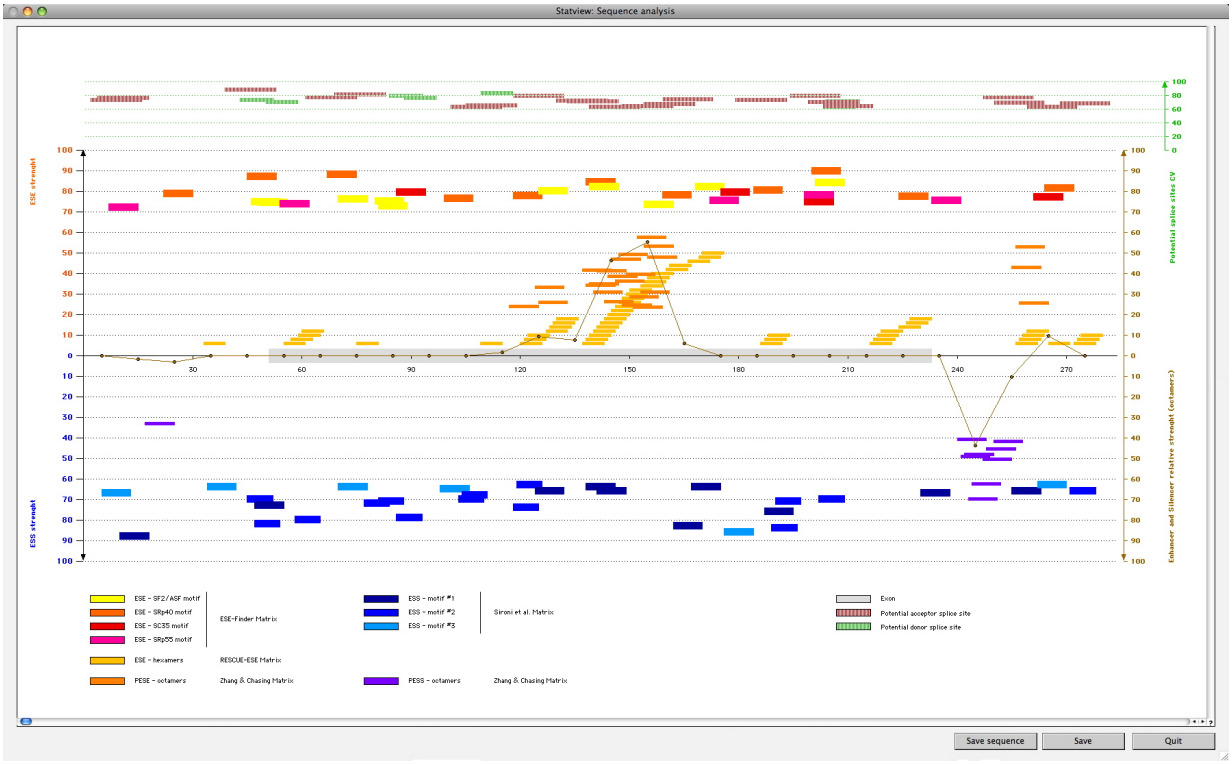


18) ESE map of exon

This function is use to predict the various auxiliary motifs (ESE) using the ESE-finder matrix. We recommend that you use the "*Sequence analysis*" function that gives more complete predictions.

19) Sequence analysis

This function is use to predict splicing signals including donor and acceptor splicesites as well as the various auxiliary motifs (ESE and ESS) using multiple matrices (for more information, view the Human Splicing Finder tool Nucleic Acid Research, 2009, April or visit the corresponding website http://www.umd.be/HSF/).

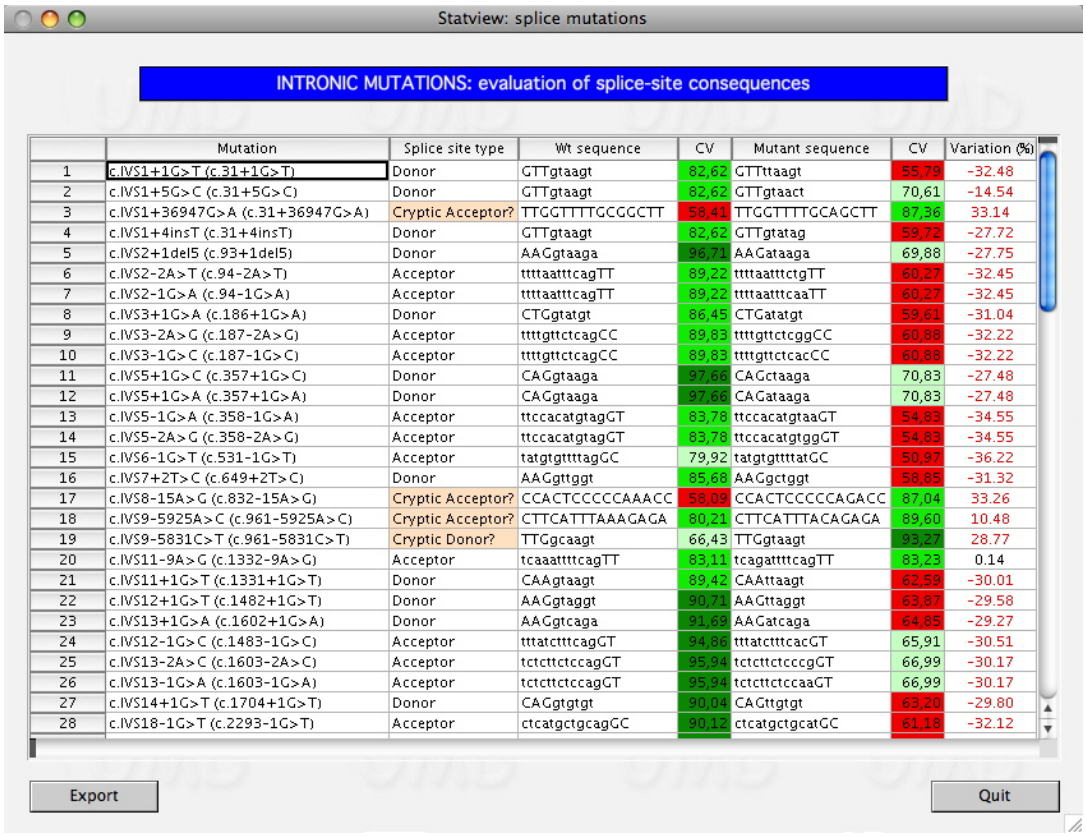


20) Splice mutations

This function is use to display the consequence of intronic mutations on splice sites and branch points.
The UMD algorithm is used for predictions and each result is displayed with a color code:

- Dark green = very strong splice site;
- Green = strong splice site;
- Light green = weaker splice site;
- Red = note a splice site.

The strength variation between the wild type and mutant potential splice sites are also given

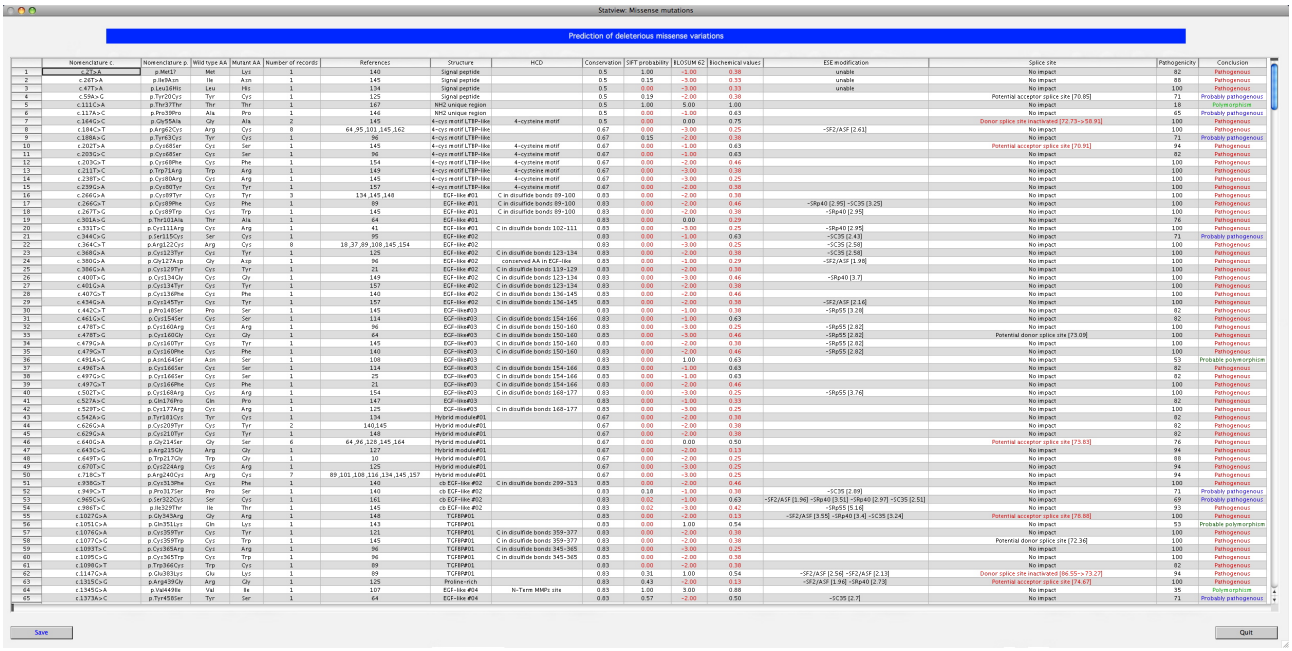


	Mutation	Splice site type	Wt sequence	CV	Mutant sequence	CV	Variation (%)
1	c.IVS1+1G>T (c.31+1G>T)	Donor	GTTgtaagt	82,62	GTTttaagt	55,79	-32.48
2	c.IVS1+5G>C (c.31+5G>C)	Donor	GTTgtaagt	82,62	GTTgtaact	70,61	-14.54
3	c.IVS1+36947G>A (c.31+36947G>A)	Cryptic Acceptor?	TTGGTTTTGCGGCTT	58,41	TTGGTTTTGCAGCTT	87,36	33.14
4	c.IVS1+4insT (c.31+4insT)	Donor	GTTgtaagt	82,62	GTTgtatag	59,72	-27.72
5	c.IVS2+1del5 (c.93+1del5)	Donor	AAGgtaaga	96,71	AAGataaga	69,88	-27.75
6	c.IVS2-2A>T (c.94-2A>T)	Acceptor	ttttaatttcagTT	89,22	ttttaatttctgTT	60,27	-32.45
7	c.IVS2-1G>A (c.94-1G>A)	Acceptor	ttttaatttcagTT	89,22	ttttaatttcaaTT	60,27	-32.45
8	c.IVS3+1G>A (c.186+1G>A)	Donor	CTGgtatgt	86,45	CTGatatgt	59,61	-31.04
9	c.IVS3-2A>G (c.187-2A>G)	Acceptor	ttttgttctcagCC	89,83	ttttgttctcggCC	60,88	-32.22
10	c.IVS3-1G>C (c.187-1G>C)	Acceptor	ttttgttctcagCC	89,83	ttttgttctcacCC	60,88	-32.22
11	c.IVS5+1G>C (c.357+1G>C)	Donor	CAGgtaaga	97,66	CAGctaaga	70,83	-27.48
12	c.IVS5+1G>A (c.357+1G>A)	Donor	CAGgtaaga	97,66	CAGataaga	70,83	-27.48
13	c.IVS5-1G>A (c.358-1G>A)	Acceptor	ttccacatgtagGT	83,78	ttccacatgtaaGT	54,83	-34.55
14	c.IVS5-2A>G (c.358-2A>G)	Acceptor	ttccacatgtagGT	83,78	ttccacatgtggGT	54,83	-34.55
15	c.IVS6-1G>T (c.531-1G>T)	Acceptor	tatgtgttttagGC	79,92	tatgtgttttatGC	50,97	-36.22
16	c.IVS7+2T>C (c.649+2T>C)	Donor	AAGgttggt	85,68	AAGgctggt	58,85	-31.32
17	c.IVS8-15A>G (c.832-15A>G)	Cryptic Acceptor?	CCACTCCCCCAAACC	58,09	CCACTCCCCCAGACC	87,04	33.26
18	c.IVS9-5925A>C (c.961-5925A>C)	Cryptic Acceptor?	CTTCATTTAAAGAGA	80,21	CTTCATTTACAGAGA	89,60	10.48
19	c.IVS9-5831C>T (c.961-5831C>T)	Cryptic Donor?	TTGgcaagt	66,43	TTGgtaagt	93,27	28.77
20	c.IVS11-9A>G (c.1332-9A>G)	Acceptor	tcaaattttcagTT	83,11	tcagattttcagTT	83,23	0.14
21	c.IVS11+1G>T (c.1331+1G>T)	Donor	CAAgtaagt	89,42	CAAttaagt	62,59	-30.01
22	c.IVS12+1G>T (c.1482+1G>T)	Donor	AAGgtaggt	90,71	AAGttaggt	63,87	-29.58
23	c.IVS13+1G>A (c.1602+1G>A)	Donor	AAGgtcaga	91,69	AAGatcaga	64,85	-29.27
24	c.IVS12-1G>C (c.1483-1G>C)	Acceptor	tttatctttcagGT	94,86	tttatctttcacGT	65,91	-30.51
25	c.IVS13-2A>C (c.1603-2A>C)	Acceptor	tctcttctccagGT	95,94	tctcttctcccgGT	66,99	-30.17
26	c.IVS13-1G>A (c.1603-1G>A)	Acceptor	tctcttctccagGT	95,94	tctcttctccaaGT	66,99	-30.17
27	c.IVS14+1G>T (c.1704+1G>T)	Donor	CAGgtgtgt	90,04	CAGttgtgt	63,20	-29.80
28	c.IVS18-1G>T (c.2293-1G>T)	Acceptor	ctcatgctgcagGC	90,12	ctcatgctgcatGC	61,18	-32.12

in %. Significant values (>10%) are indicated in red.
Note that deep intronic mutations from this example are predicted to activate acceptor or donor cryptic splice sites. All corresponding predictions have been proven experimentally. For more information UMD-DMD database, Human Mutation.

21) Missense mutations

We have created the UMD-Predictor® tool to predict the pathogenicity of substitutions localized in the coding sequence. To have more information about how it works, read the UMD-Predictor paper, Human Mutation.

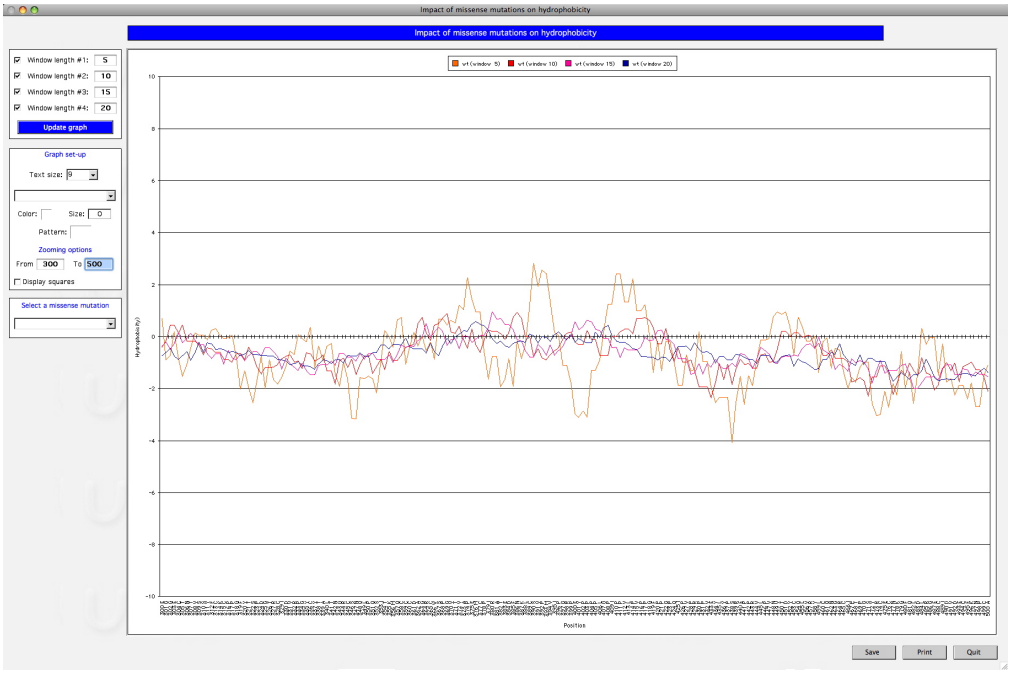


22) All missense variations

This function allows the prediction of pathogenicity of all possible substitutions. It usually takes some time to run and display. The result can be saved for further uses.

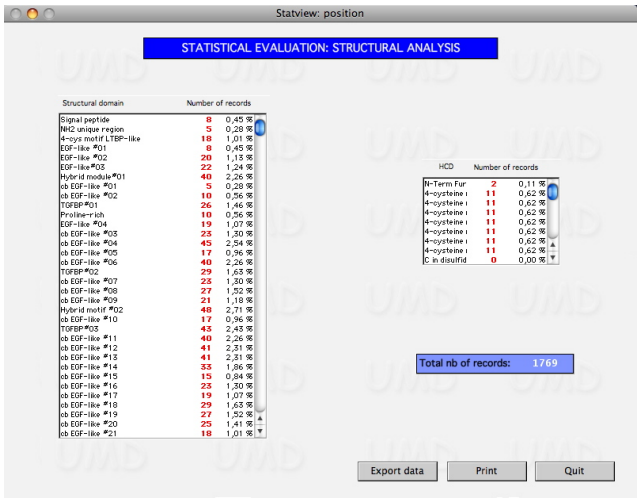
23) Missenses and hydrophobicity

This function is primarily dedicated to transmembrane proteins. The user can select different window sizes to evaluate the hydrophobicity. It is also possible to zoom on specific regions of the protein and finally evaluate the impact of a missense mutation of the hydrophobicity profiles.



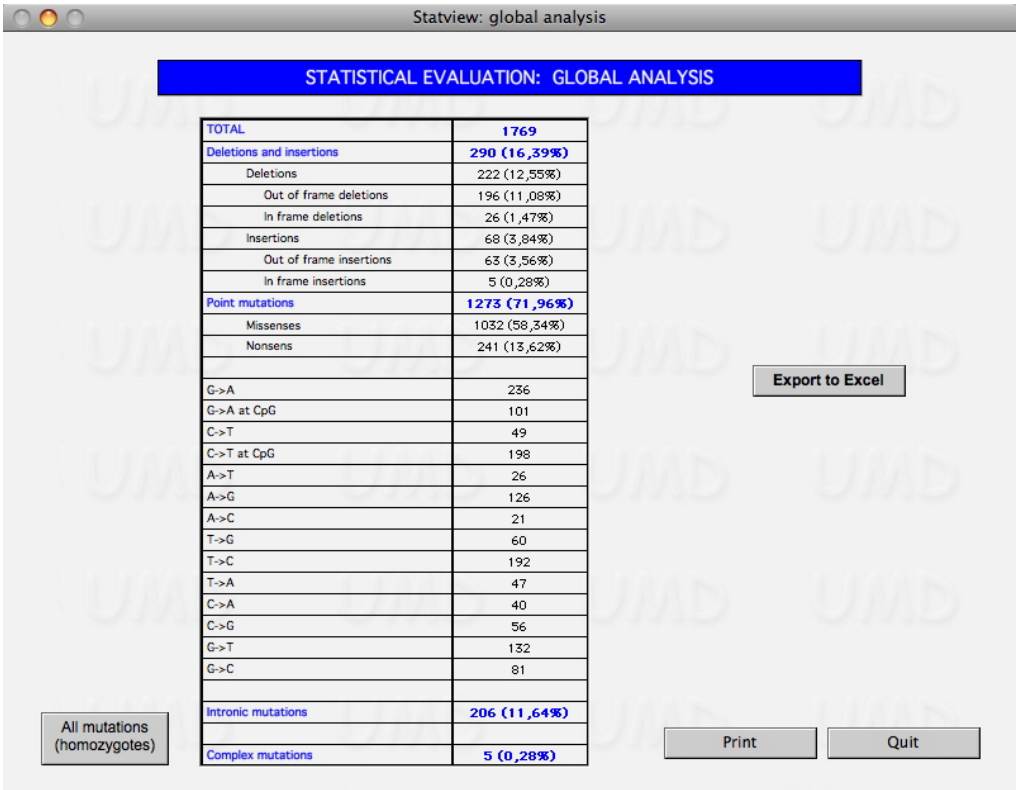
24) Structure

This function is used to list the various structural domains of the protein where are localized the mutations. The same analyze is also performed at the HCD level.

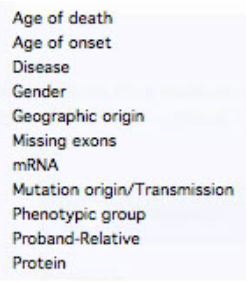


25) Global analysis

This function was designed to give a summary of mutations types related to a group of records. The "*All mutations (homozygous)*" button can be used to avoid bias when genes with recessive transmission are involved. In this situation homozygous mutations are counted twice. The results can be exported as a text-tab delimited file using the "*Export to Excel*" button.



26) Free Graph

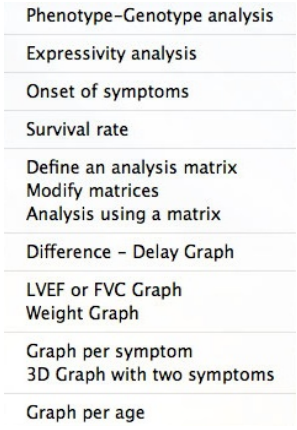


This analysis can be performed either on all patients or on a pre-selection of records. The field for the analysis can be chosen in a given list (depending of the gene. For a specific design, please contact us).
A graphical display of the field values is displayed. The user can choose the chart type.
This graph can be saved as picture using the "*Save*" button or printed using the "*Print*" button.

27) Specific tools

This function is activated on a collaborative base when a curator needs specific tools. For example tools related to exon skipping for the DMD gene...
If you need specific tools that are not yet available in the UMD software, please contact us.

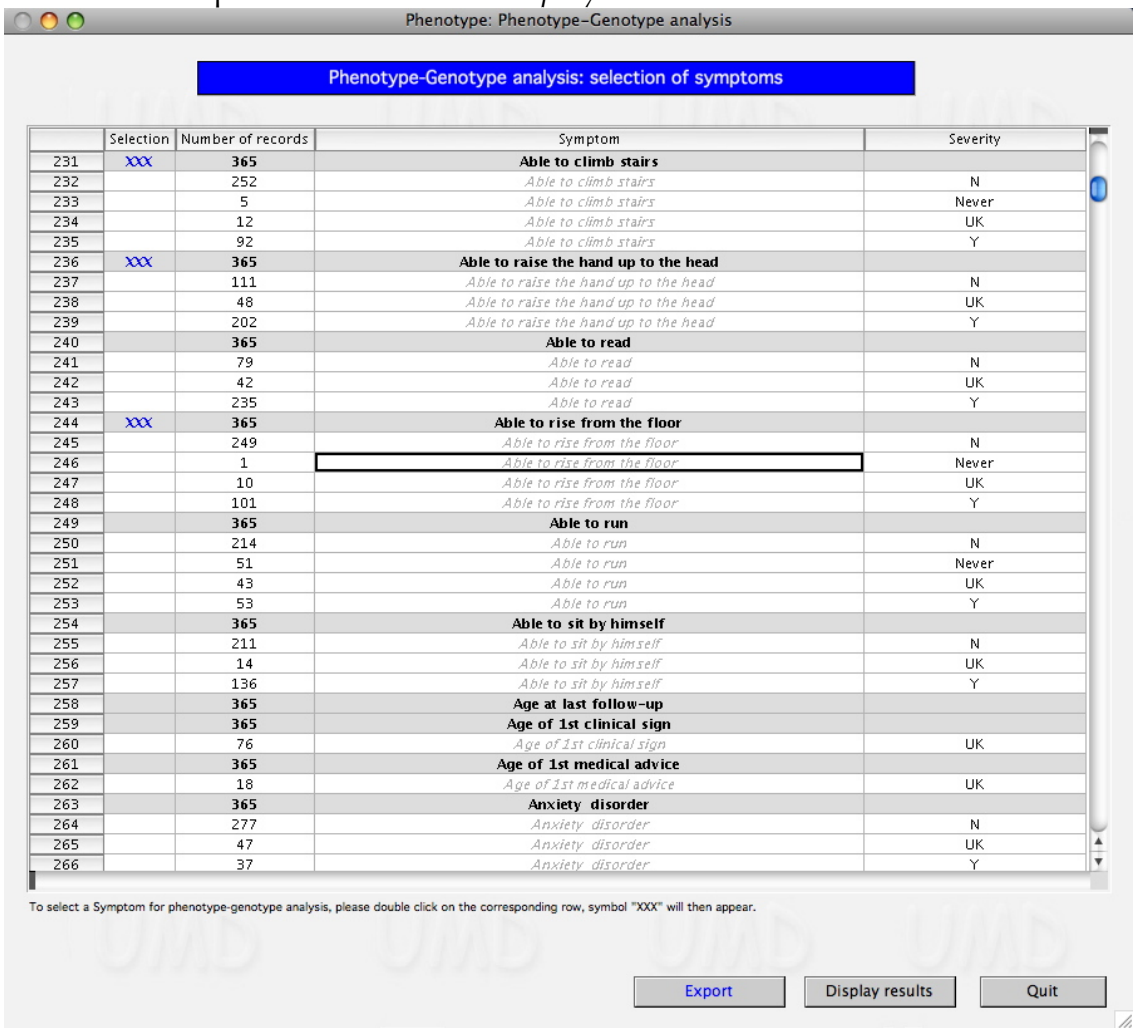
VIII-g- The "Phenotype" menu



This menu contains 13 functions, which are related to the clinical data analysis either using statistical or graphical presentations. Note that because clinical data are stored in an alpha-numeric field (severity), specific parameters should be used for proper analysis (cf. VIII-a-30 31 and 32).

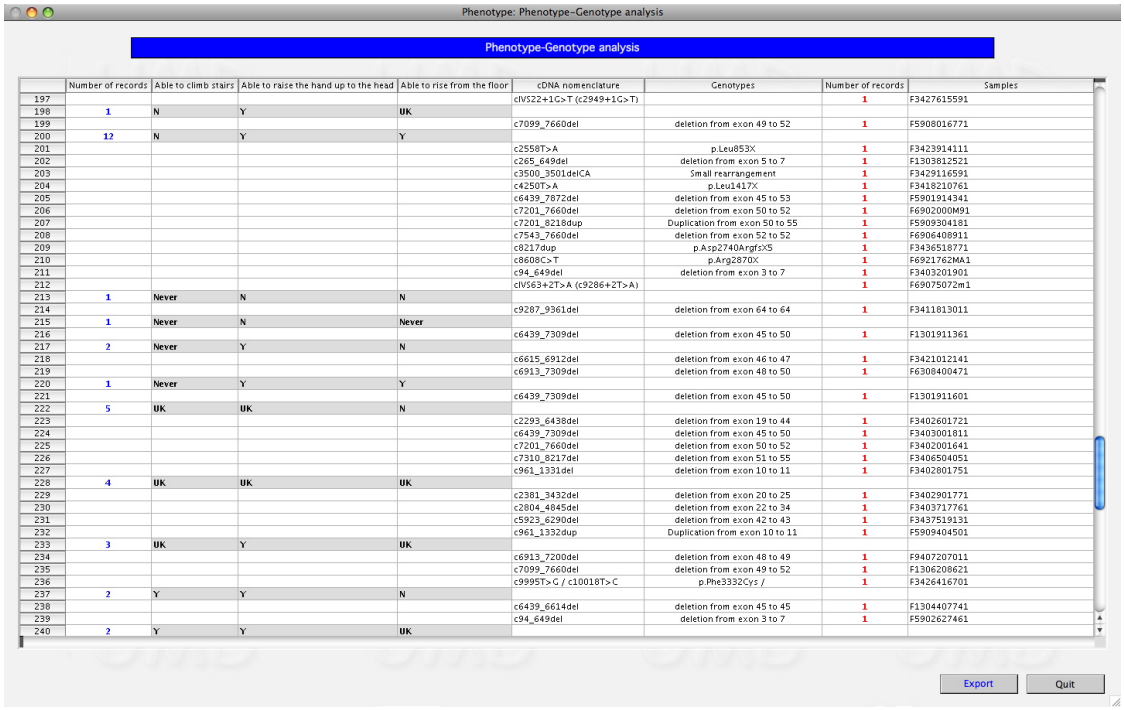
1) Phenotype-genotype analysis

Be sure that you first have used the "*Update clinical data*" function prior to use this function. A summary table is then displayed with, for each symptom, the corresponding severities and the number of patients.
Each symptom is displayed as a light grey row and the corresponding severities are listed below. The user can select a symptom by a double-click either on the symptom row or on any row corresponding to an associated severity. A triple star symbol (***) will then appear in the selection column. A new double-click on the same lines will deselect the symptom. The analysis can then be performed with the "*Display results*" button.



	Selection	Number of records	Symptom	Severity
231	XXX	365	**Able to climb stairs**	
232		252	*Able to climb stairs*	N
233		5	*Able to climb stairs*	Never
234		12	*Able to climb stairs*	UK
235		92	*Able to climb stairs*	Y
236	XXX	365	**Able to raise the hand up to the head**	
237		111	*Able to raise the hand up to the head*	N
238		48	*Able to raise the hand up to the head*	UK
239		202	*Able to raise the hand up to the head*	Y
240		365	**Able to read**	
241		79	*Able to read*	N
242		42	*Able to read*	UK
243		235	*Able to read*	Y
244	XXX	365	**Able to rise from the floor**	
245		249	*Able to rise from the floor*	N
246		1	*Able to rise from the floor*	Never
247		10	*Able to rise from the floor*	UK
248		101	*Able to rise from the floor*	Y
249		365	**Able to run**	
250		214	*Able to run*	N
251		51	*Able to run*	Never
252		43	*Able to run*	UK
253		53	*Able to run*	Y
254		365	**Able to sit by himself**	
255		211	*Able to sit by himself*	N
256		14	*Able to sit by himself*	UK
257		136	*Able to sit by himself*	Y
258		365	**Age at last follow-up**	
259		365	**Age of 1st clinical sign**	
260		76	*Age of 1st clinical sign*	UK
261		365	**Age of 1st medical advice**	
262		18	*Age of 1st medical advice*	UK
263		365	**Anxiety disorder**	
264		277	*Anxiety disorder*	N
265		47	*Anxiety disorder*	UK
266		37	*Anxiety disorder*	Y

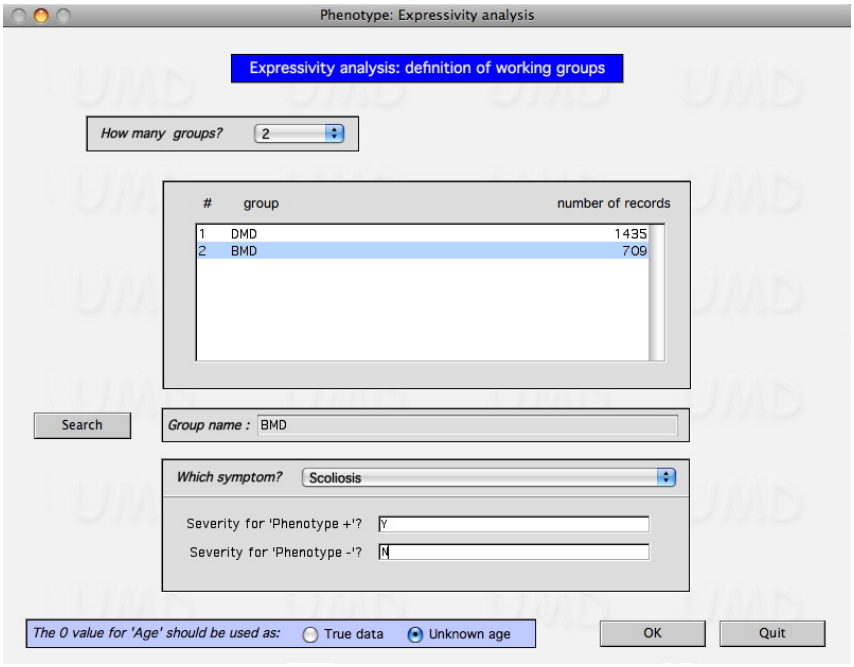
The various combinations of symptoms and severities are then computed and the number of patients and associated genotypes are listed.



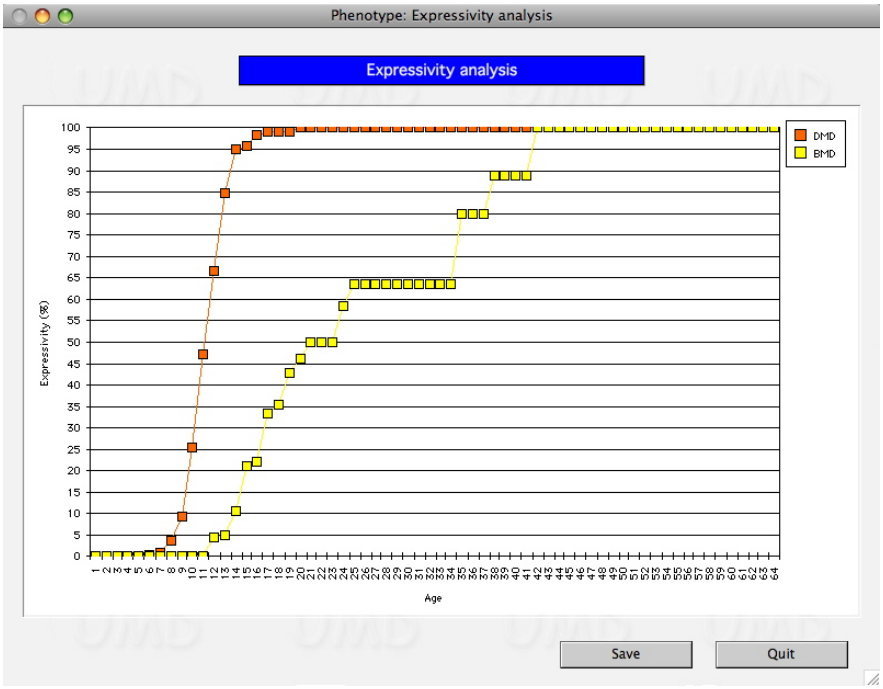
In addition, the "Samples" column give an easy access to the corresponding patients. The results can be exported via the "Export" button, while the "Quit" button returns the user to the full list of symptoms and associated severities in order to perform another analysis.

2) Expressivity analysis

This function is used to study the expressivity of a specific symptom within one or multiple



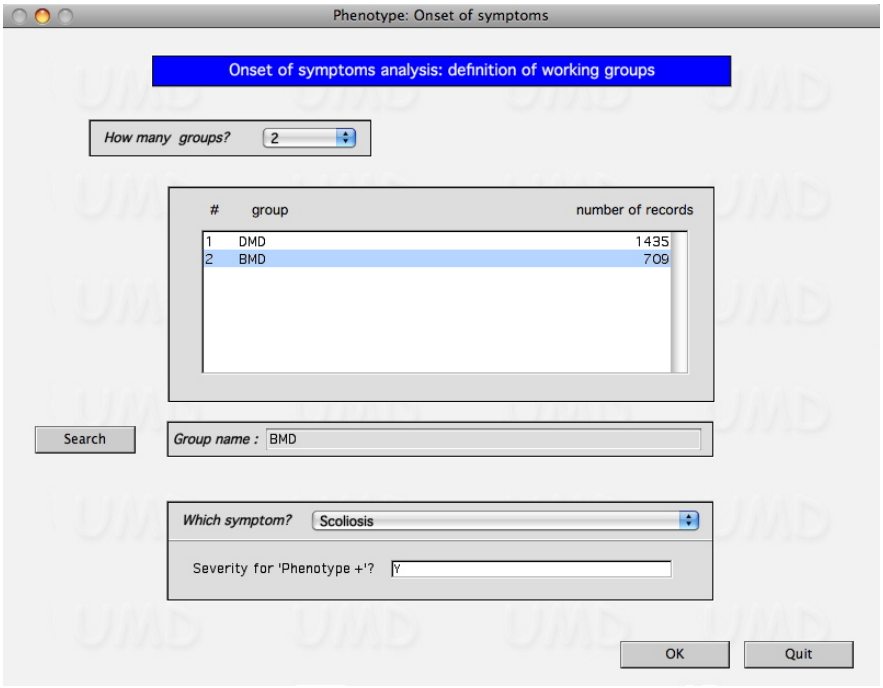
groups (up to 8) of records defined by the user.
The user then selects a symptom and defines the severities associated to the "+" and "-" phenotype. The value 0 can either be used as a true data or as an unknown age as the default value for the "Age" field is 0 even if no data was available.



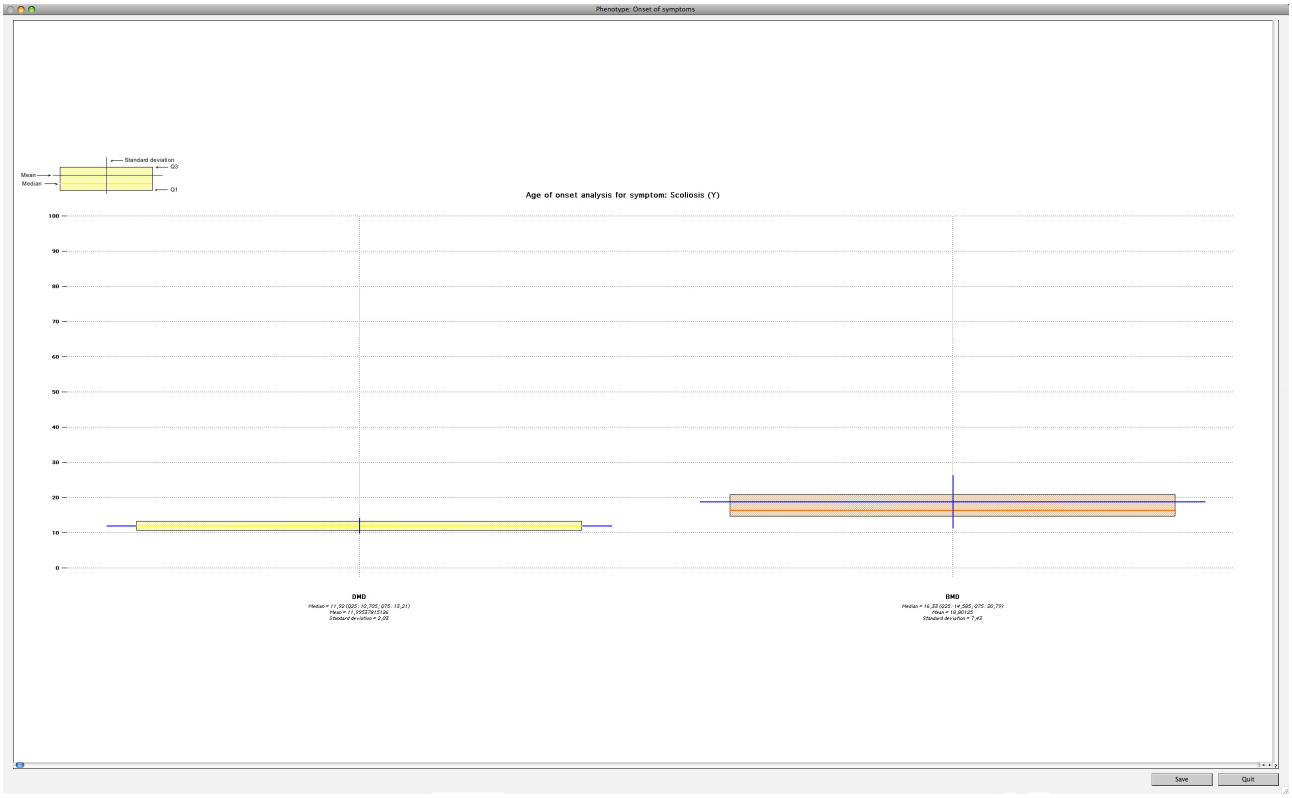
A graph with the expressivity (percentage of patients for which a data is available at a specific age and present the "+" phenotype) is then shown. The "*Quit*" button returns to the previous screen to perform other analysis.

3) Onset of symptoms

This function is used to study the age of onset of a specific symptom. As for the previous analysis, the user can define various groups of records and choose the symptom and the



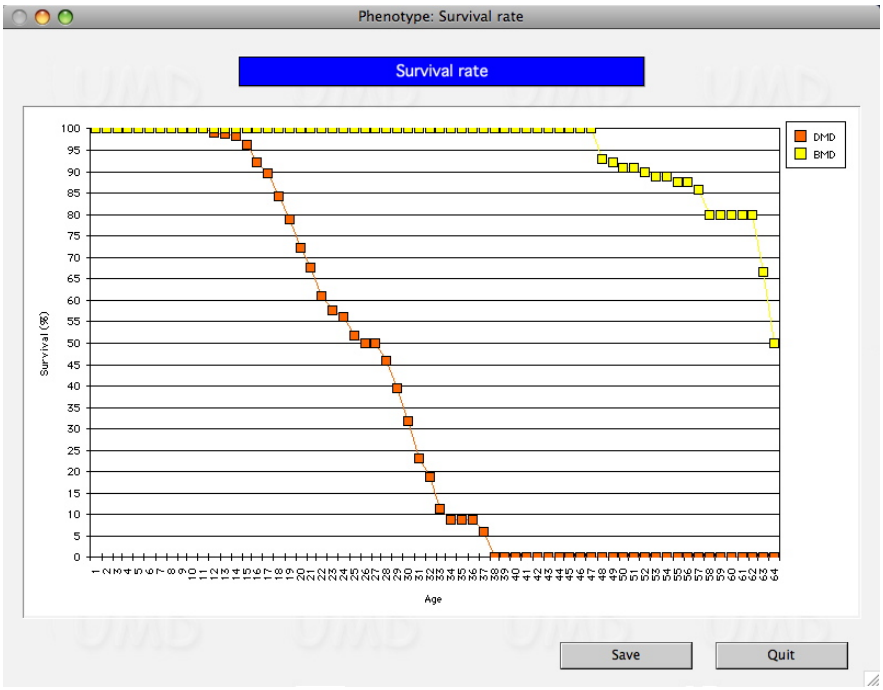
severity associated with the "+" phenotype.
When the parameters are validated, a click on the "OK" button will open the result screen.



The following data are available for each group of records: Mean, median, standard deviation, first and third quartiles.

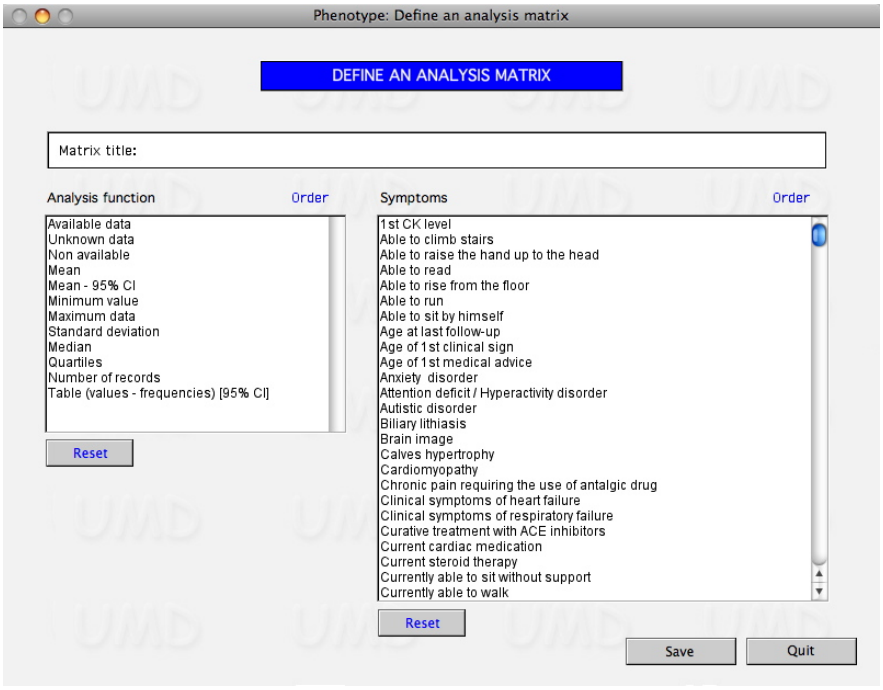
4) Survival rate

The user first defines the various groups of records that should be used for this analysis (up to 8). The UMD® software will then analyze the data using the "*Age of death*" field and define the percentage of patients still alive at a specific age.



5) Define an analysis matrix

The analysis matrices are used to perform statistical analysis from clinical data. The user can create a virtually unlimited number of matrices. For each, he needs to select one or more "*Analysis functions*" from the list (left) and select one or more symptoms (right). When the



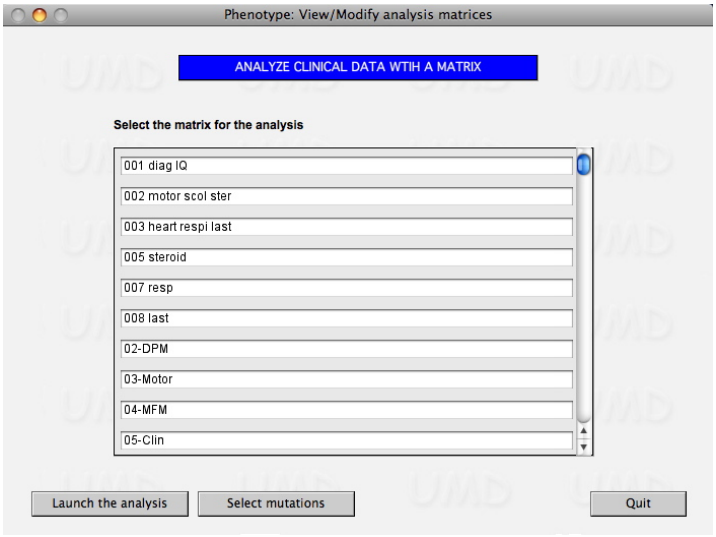
selections are completed, a unique matrix title should also be given (top).

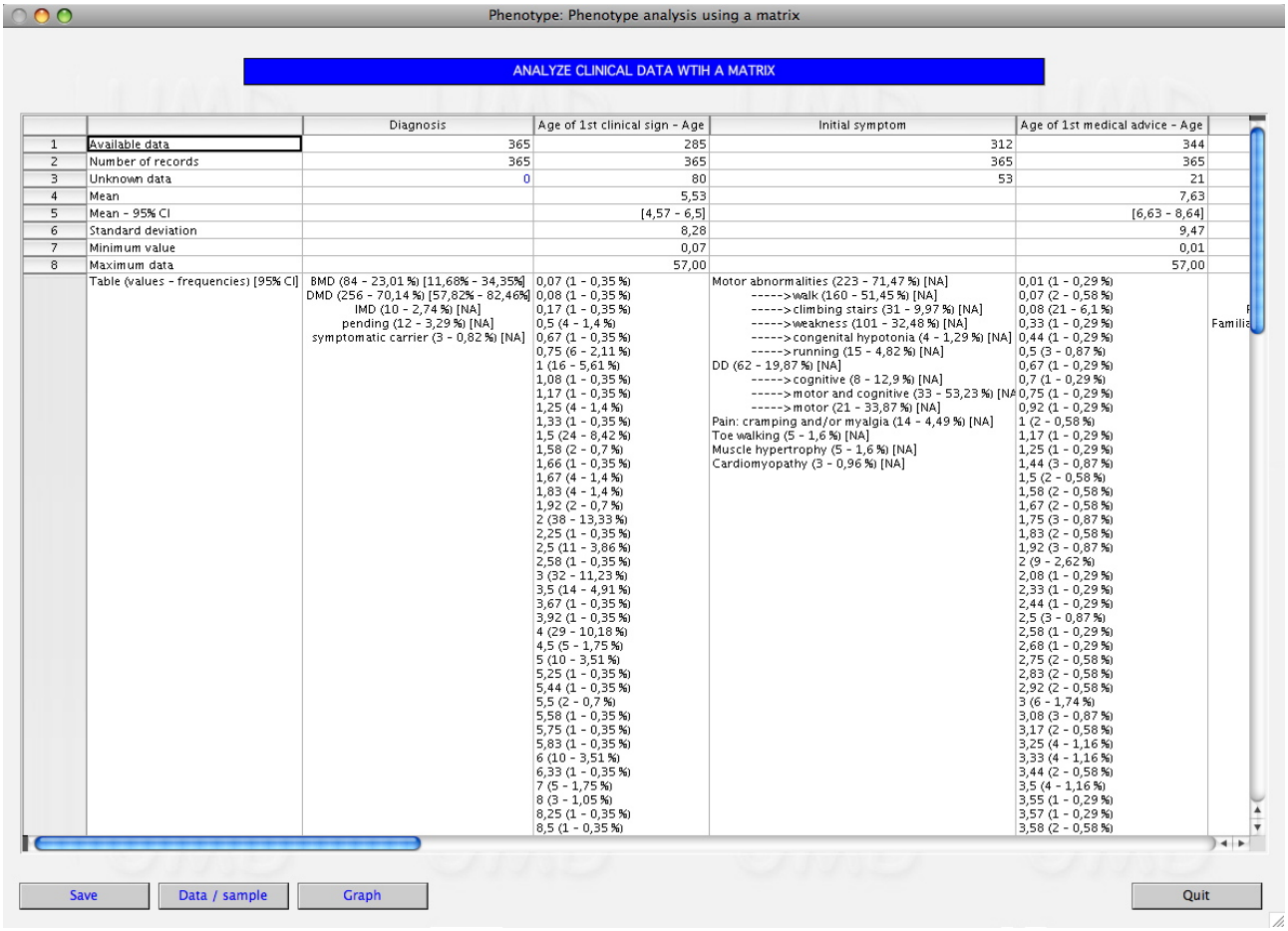
6) Modify matrices

This function lists all available matrices and allow their deletion or modification.

7) Analysis using a Matrix

The user can use a specific matrix from the list on all or a subset of clinical records. The "*Select mutations*" button gives access to the search interface in order to extract a group of records using any of the data fields (for more information about record extraction, cf. chapter VII).
A simple click on the matrices list will select the appropriate matrix and the "*Launch the analysis*" button will perform the analysis.



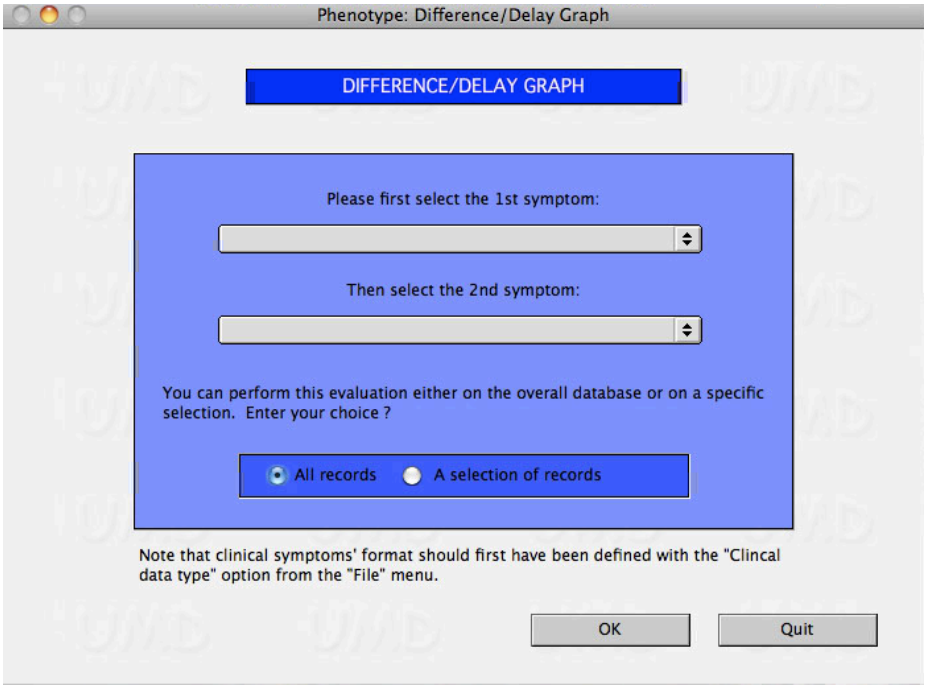


		Diagnosis	Age of 1st clinical sign - Age	Initial symptom	Age of 1st medical advice - Age	
1	Available data	365	285	312	344	
2	Number of records	365	365	365	365	
3	Unknown data	0	80	53	21	
4	Mean		5,53		7,63	
5	Mean - 95% CI		[4,57 - 6,5]		[6,63 - 8,64]	
6	Standard deviation		8,28		9,47	
7	Minimum value		0,07		0,01	
8	Maximum data		57,00		57,00	
	Table (values - frequencies) [95% CI]	BMD (84 - 23,01 %) [11,68% - 34,35%] DMD (256 - 70,14 %) [57,82% - 82,46%] IMD (10 - 2,74 %) [NA] pending (12 - 3,29 %) [NA] symptomatic carrier (3 - 0,82 %) [NA]	0,07 (1 - 0,35 %) 0,08 (1 - 0,35 %) 0,17 (1 - 0,35 %) 0,5 (4 - 1,4 %) 0,67 (1 - 0,35 %) 0,75 (6 - 2,11 %) 1 (16 - 5,61 %) 1,08 (1 - 0,35 %) 1,17 (1 - 0,35 %) 1,25 (4 - 1,4 %) 1,33 (1 - 0,35 %) 1,5 (24 - 8,42 %) 1,58 (2 - 0,7 %) 1,66 (1 - 0,35 %) 1,67 (4 - 1,4 %) 1,83 (4 - 1,4 %) 1,92 (2 - 0,7 %) 2 (38 - 13,33 %) 2,25 (1 - 0,35 %) 2,5 (11 - 3,86 %) 2,58 (1 - 0,35 %) 3 (32 - 11,23 %) 3,5 (14 - 4,91 %) 3,67 (1 - 0,35 %) 3,92 (1 - 0,35 %) 4 (29 - 10,18 %) 4,5 (5 - 1,75 %) 5 (10 - 3,51 %) 5,25 (1 - 0,35 %) 5,44 (1 - 0,35 %) 5,5 (2 - 0,7 %) 5,58 (1 - 0,35 %) 5,75 (1 - 0,35 %) 5,83 (1 - 0,35 %) 6 (10 - 3,51 %) 6,33 (1 - 0,35 %) 7 (5 - 1,75 %) 8 (3 - 1,05 %) 8,25 (1 - 0,35 %) 8,5 (1 - 0,35 %)	Motor abnormalities (223 - 71,47 %) [NA] ----->walk (160 - 51,45 %) [NA] ----->climbing stairs (31 - 9,97 %) [NA] ----->weakness (101 - 32,48 %) [NA] ----->congenital hypotonia (4 - 1,29 %) [NA] ----->running (15 - 4,82 %) [NA] DD (62 - 19,87 %) [NA] ----->cognitive (8 - 12,9 %) [NA] ----->motor and cognitive (33 - 53,23 %) [NA] ----->motor (21 - 33,87 %) [NA] Pain: cramping and/or myalgia (14 - 4,49 %) [NA] Toe walking (5 - 1,6 %) [NA] Muscle hypertrophy (5 - 1,6 %) [NA] Cardiomyopathy (3 - 0,96 %) [NA]	0,01 (1 - 0,29 %) 0,07 (2 - 0,58 %) 0,08 (21 - 6,1 %) 0,33 (1 - 0,29 %) 0,44 (1 - 0,29 %) 0,5 (3 - 0,87 %) 0,67 (1 - 0,29 %) 0,7 (1 - 0,29 %) 0,75 (1 - 0,29 %) 0,92 (1 - 0,29 %) 1 (2 - 0,58 %) 1,17 (1 - 0,29 %) 1,25 (1 - 0,29 %) 1,44 (3 - 0,87 %) 1,5 (2 - 0,58 %) 1,58 (2 - 0,58 %) 1,67 (2 - 0,58 %) 1,75 (3 - 0,87 %) 1,83 (2 - 0,58 %) 1,92 (3 - 0,87 %) 2 (9 - 2,62 %) 2,08 (1 - 0,29 %) 2,33 (1 - 0,29 %) 2,44 (1 - 0,29 %) 2,5 (3 - 0,87 %) 2,58 (1 - 0,29 %) 2,68 (1 - 0,29 %) 2,75 (2 - 0,58 %) 2,83 (2 - 0,58 %) 2,92 (2 - 0,58 %) 3 (6 - 1,74 %) 3,08 (3 - 0,87 %) 3,17 (2 - 0,58 %) 3,25 (4 - 1,16 %) 3,33 (4 - 1,16 %) 3,44 (2 - 0,58 %) 3,5 (4 - 1,16 %) 3,55 (1 - 0,29 %) 3,57 (1 - 0,29 %) 3,58 (2 - 0,58 %)	F Familia

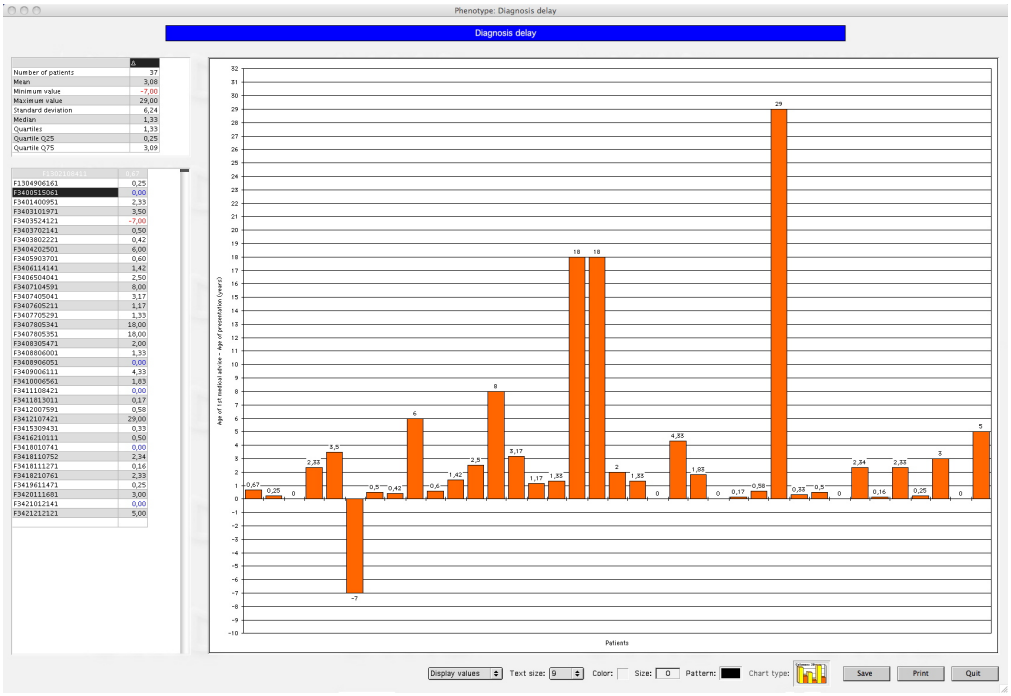
In this example 9 "analysis functions" have been selected and applied to various symptoms. Note that when severities are texts (Diagnosis, Initial symptom) only few functions could be applied. When numeric values or age are used (Age of first clinical sign, Age of first medical advice), all functions are used. Note that for all types of data are given: the various values, the number of occurrence, their frequency and the 95% confidence interval.

8) Difference - Delay Graph

This function was designed to evaluate the diagnosis delay, but can also be used for other purposes (differences between 2 dates or numerical values). It displays the difference between two “clinical items” chosen by the user for all records or for a selection of records. The chart type can be changed.
The user needs to select two symptoms: for a typical diagnostic delay graph, the first symptom is the “age of first medical advice” and the second symptom is the “age of presentation”.



The Y-axis represents the difference between the 2 values. In addition are presented the: mean, minimum and maximum values, standard deviation, median, quartiles, quartile Q25 and quartile Q75. The X-axis represents each individual.

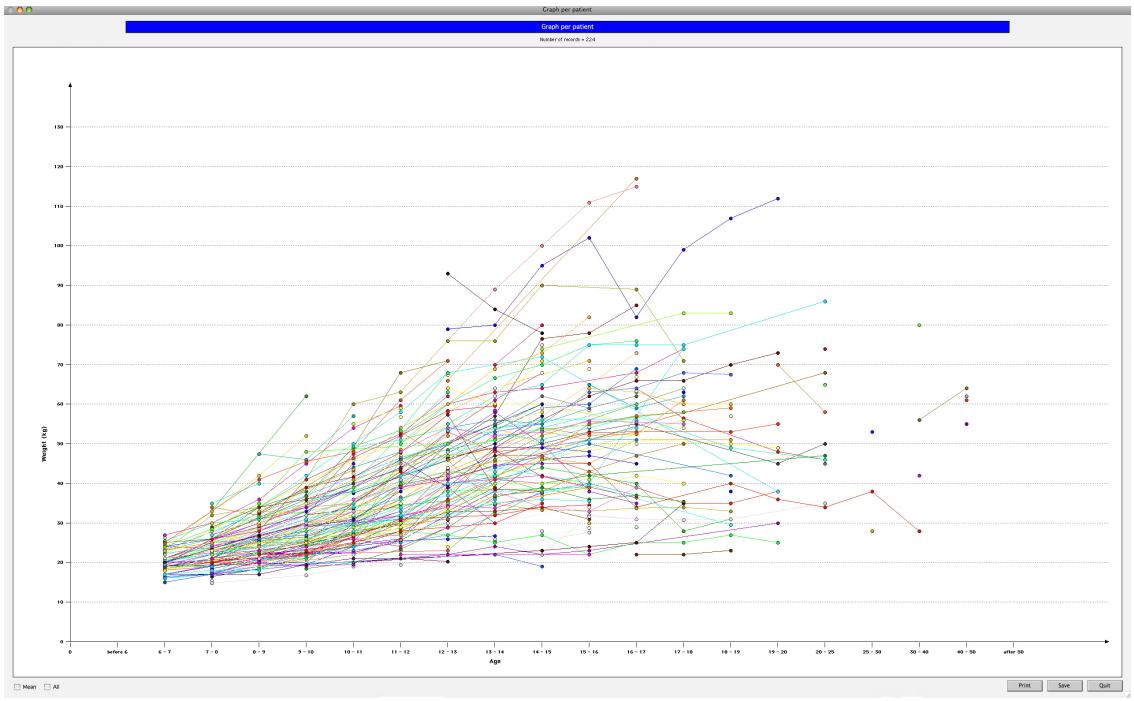


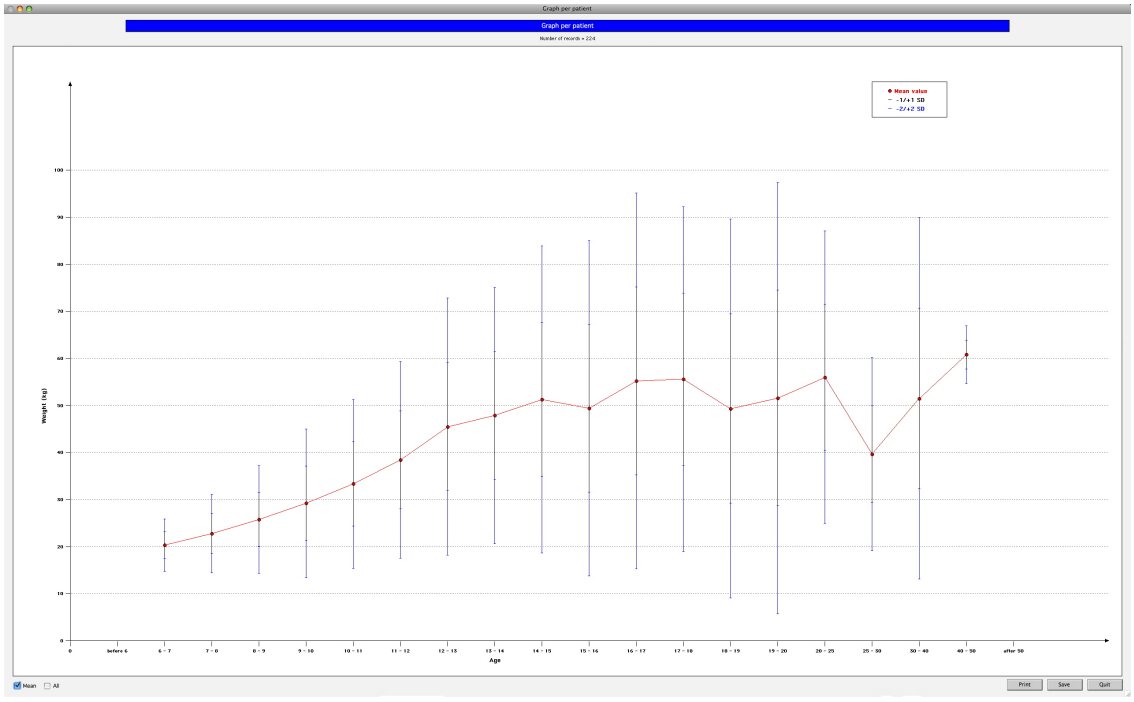
9) LVEF or FVC graph

These functions are only available for few curators. If you want to activate these functions, please contact us for more details.

10) Weight graph

This analysis can be performed either on one or all patients. A pre-selection of records can also be performed.
A graphical display of the weight evolution is displayed. The user can choose to combine all data and visualize mean values and standard deviations in a second graph.

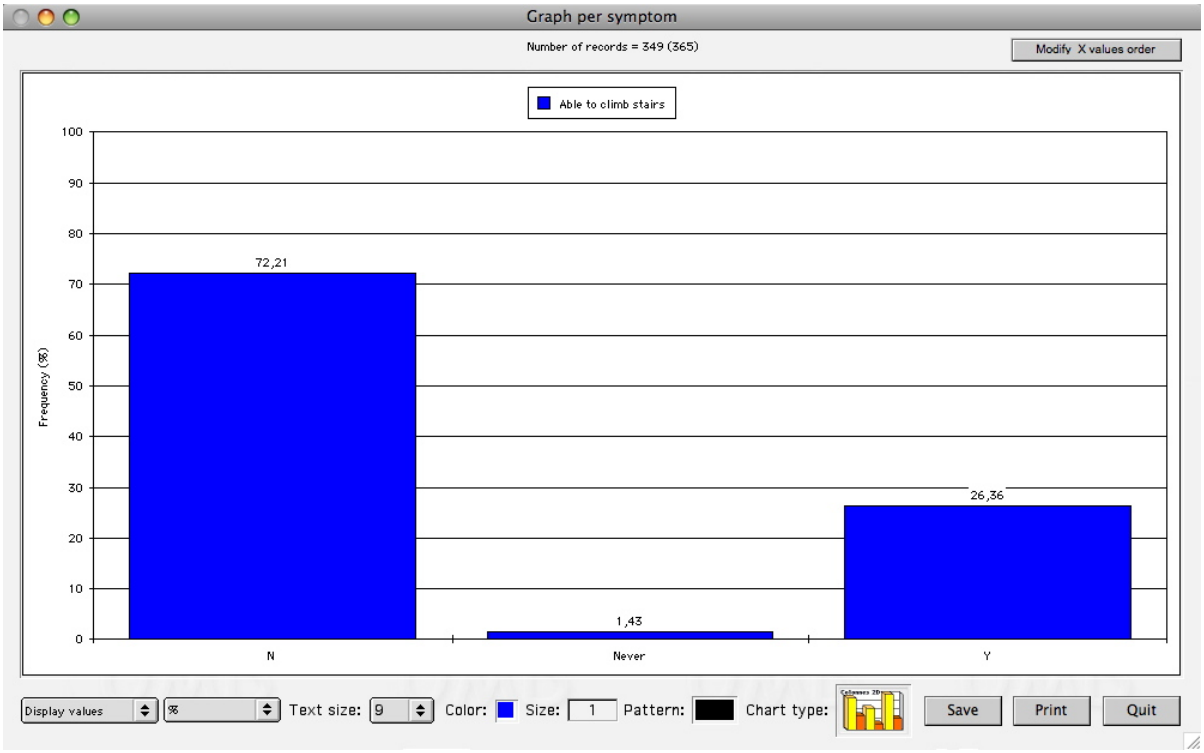




Both graphs can be saved as pictures using the "*Save*" buttons or printed using the "*Print*" buttons.

11) Graph per symptom

This analysis can be performed on any symptom from the clinical data list. The user can



choose to perform it on all or only a subset of records.
The user can modify the graphics using the following tools:

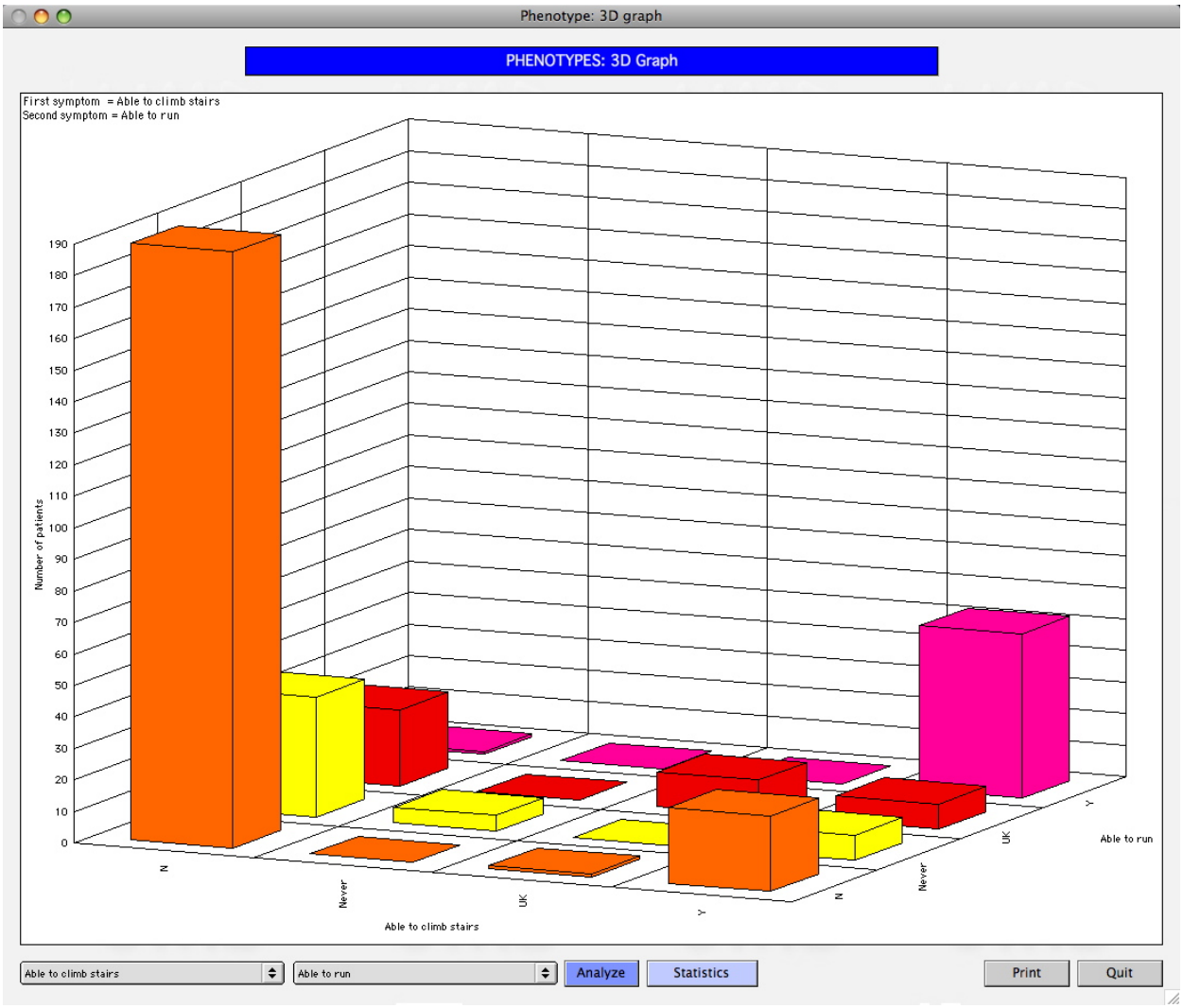
- The "Display values" pop-up menu allows to hide or display the values;
- The "%" pop-up menu allows to display either the number of patients or the% values;
- The "Text size" to change the text size used in the graph;
- The "Color" can be changed using the 256 colors pallet;
- The "size" of border lines and the graph "Pattern";
- The "Chart type" to modify the graph type.

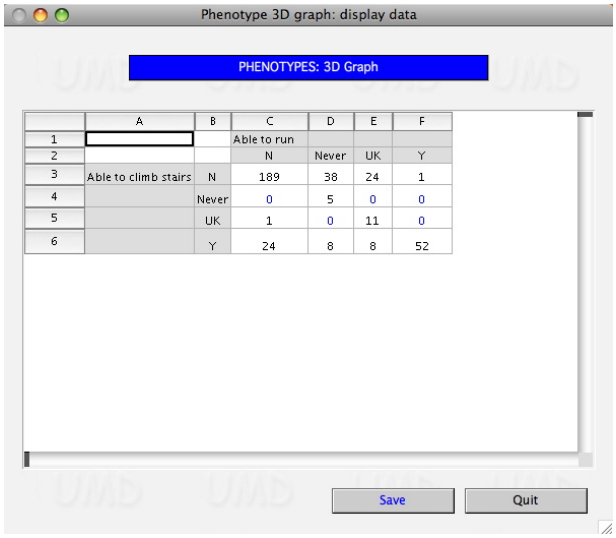
In addition, the "*Modify X values order*" button can be used to modify the order of severities within the graph.

12) 3D graph with two symptoms

This function is used to combine 2 symptoms and display the various combinations of associated severities. This analysis can be performed on a selection of records such as a specific disease or phenotype.
The user first need to select 2 symptoms, he will be afterwards able to modify these symptoms directly from the graph.
The 3D graph vie can be modified using the left, right, up and down arrows.

Two pop-up menus allows the selection of different symptoms and the "Analyze" button will allow the update of the graph using these new parameters.
Finally, the "Statistics" button will give access to a table with the numerical values used to draw the graph. These data can be exported.



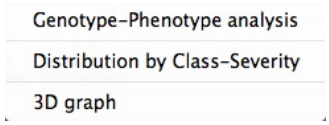


13) Graph per age

This analysis can be performed on any symptom from the clinical data list. The user can choose to perform it on all or only a subset of records.
Two graphs will be displayed, the first one is similar to the one presented for the "*Graph per symptom*" function, while the second one has the same display than the "*Onset of symptoms*" function.

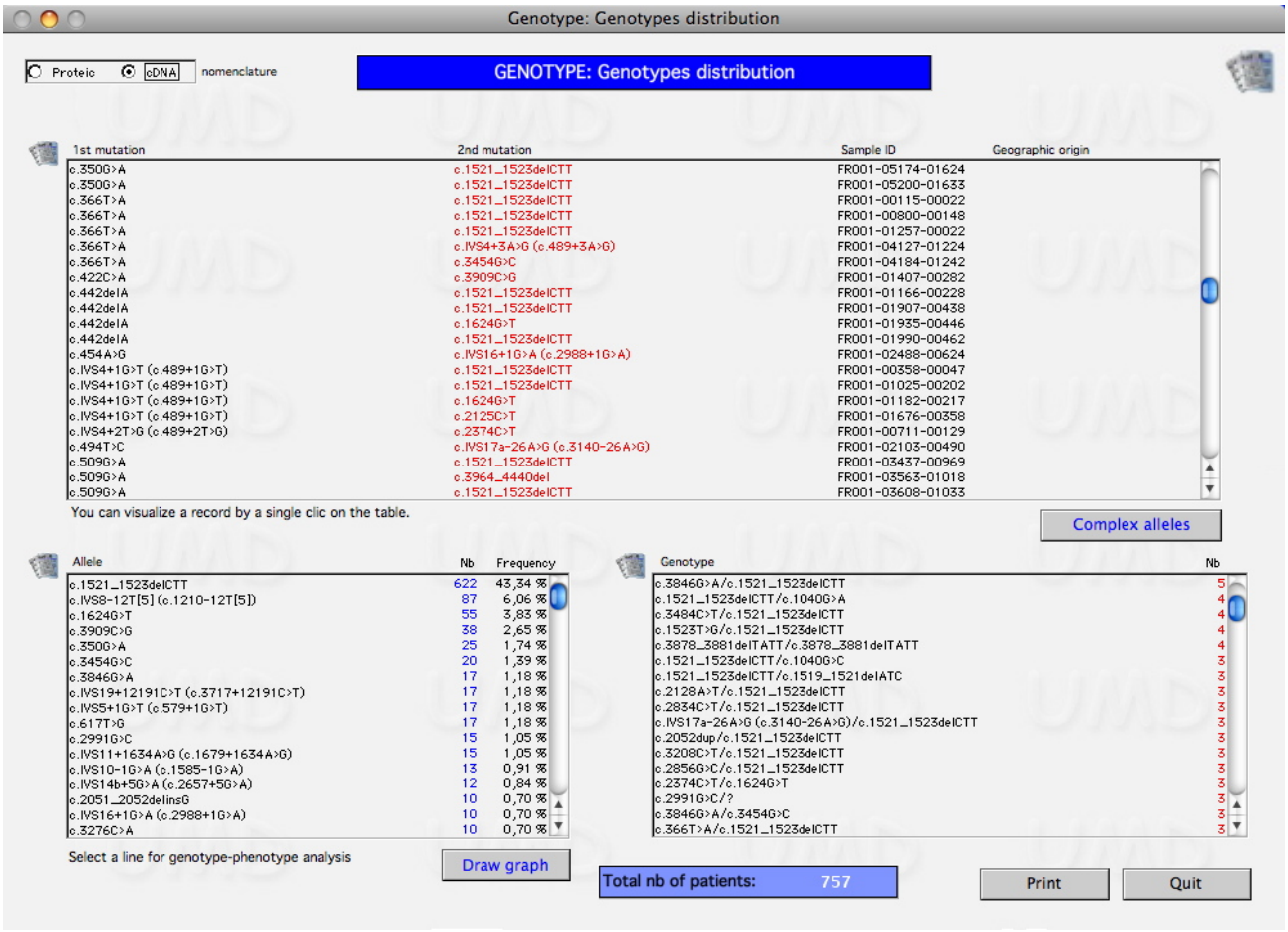
VIII-h- The "Genotype" menu

This menu contains 3 functions related to genotypes.



1) Genotype-Phenotype analysis

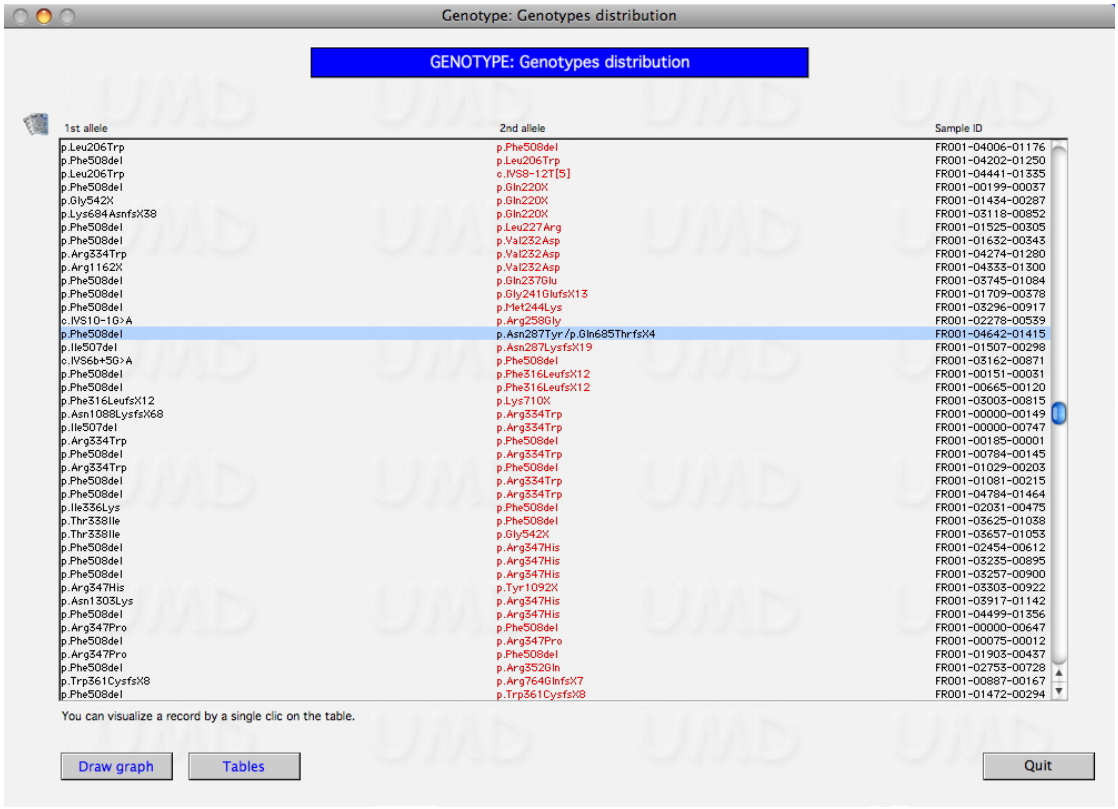
This function is used to display mutations (one or two alleles per patient), alleles and genotypes frequencies.



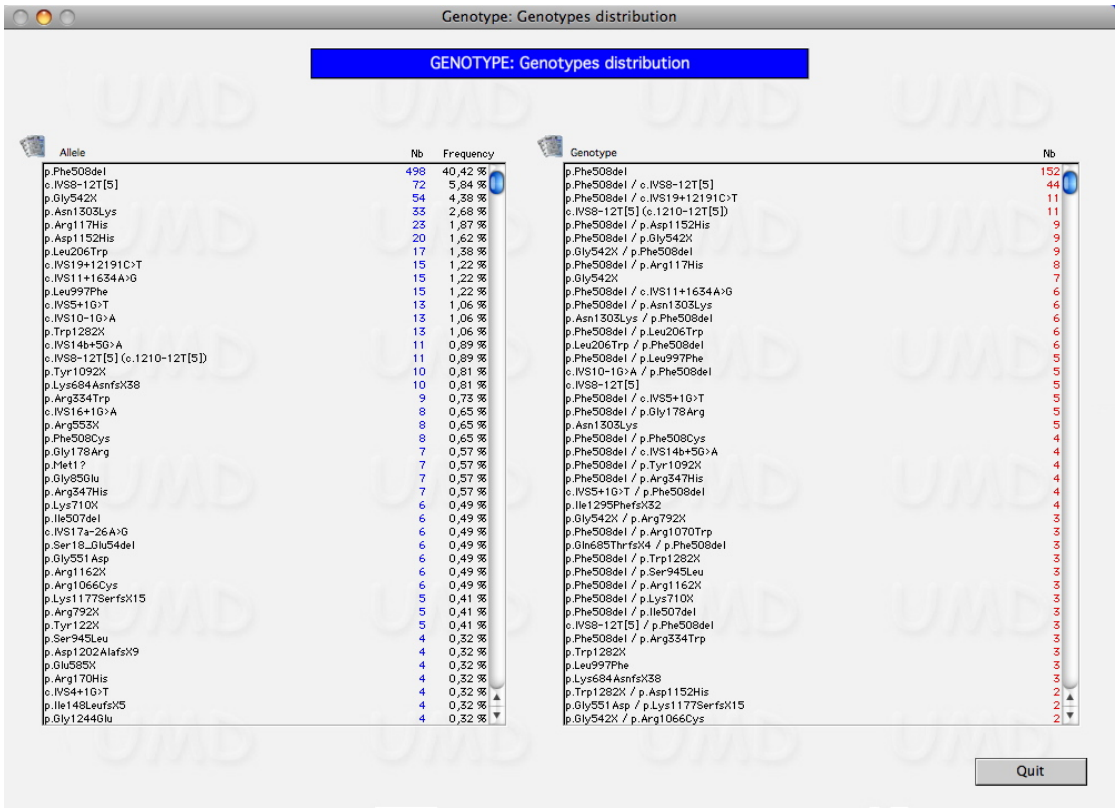
A simple click on the table gives access to the corresponding patient records.
To save the corresponding data of each table as a text-tab delimited file, click on the symbol.

The "*Draw graph*" button opens a graph that displays the various number of genotypes (see below) or alleles (from the upper screen).

A click on the "*Complex allele*" button opens a new window to display the complete genotype and therefore to complex alleles (see below). A simple click on the table gives access to the patient records.



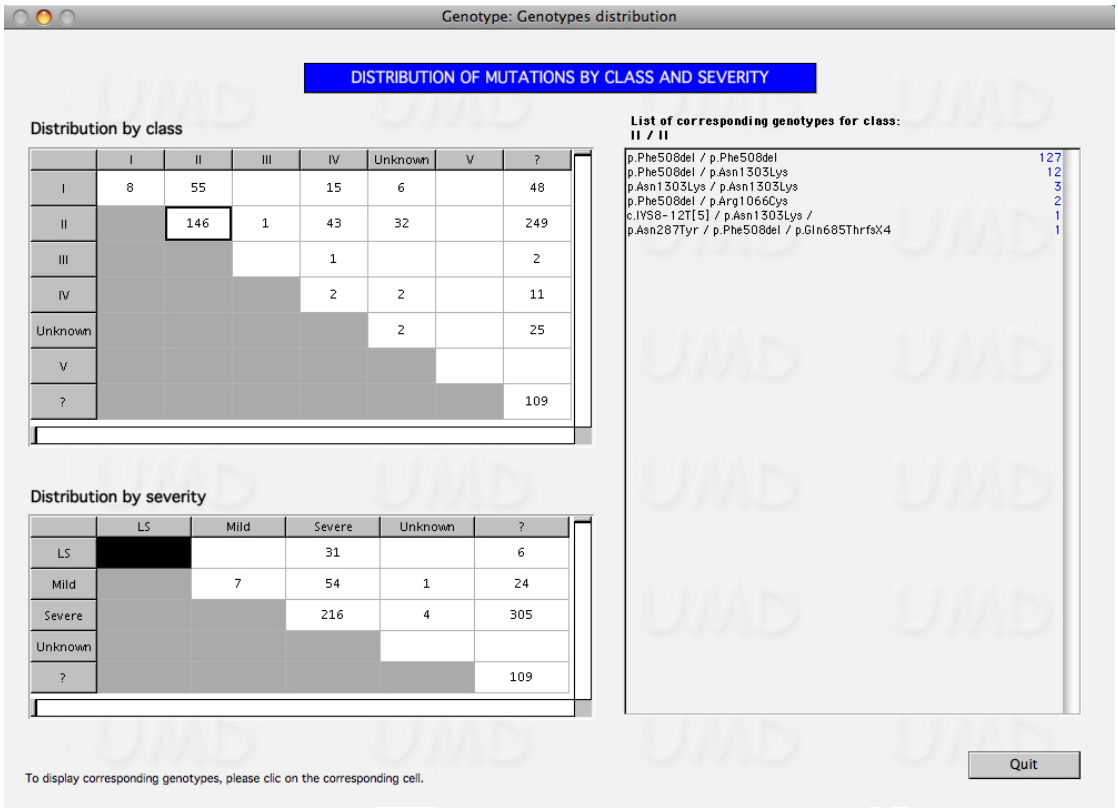
Note that the "FR001-04642-01415" record harbors two alleles in the CFTR gene: The frequent p.Phe508del mutation and a complex allele p.Asn287Tyr/p.Gln6851ThrfsX4. A click on the "Tables" button opens the following windosw:



All data can be exported via the corresponding symbol.

2) Distribution by Class-Severity

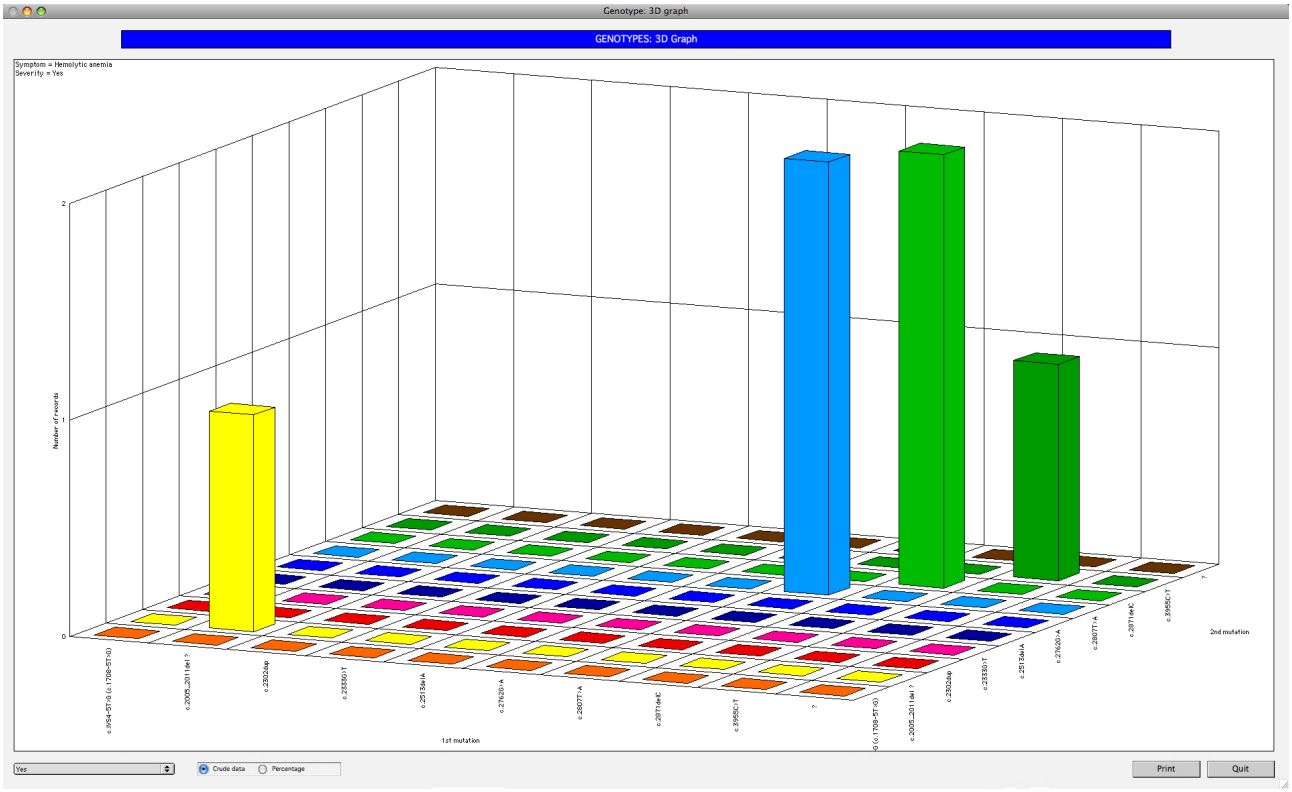
In the Variation table (see chapter VIII-l), each mutation can be associated to a mutation class and a mutation severity. These two annotations are used in this function for recessive



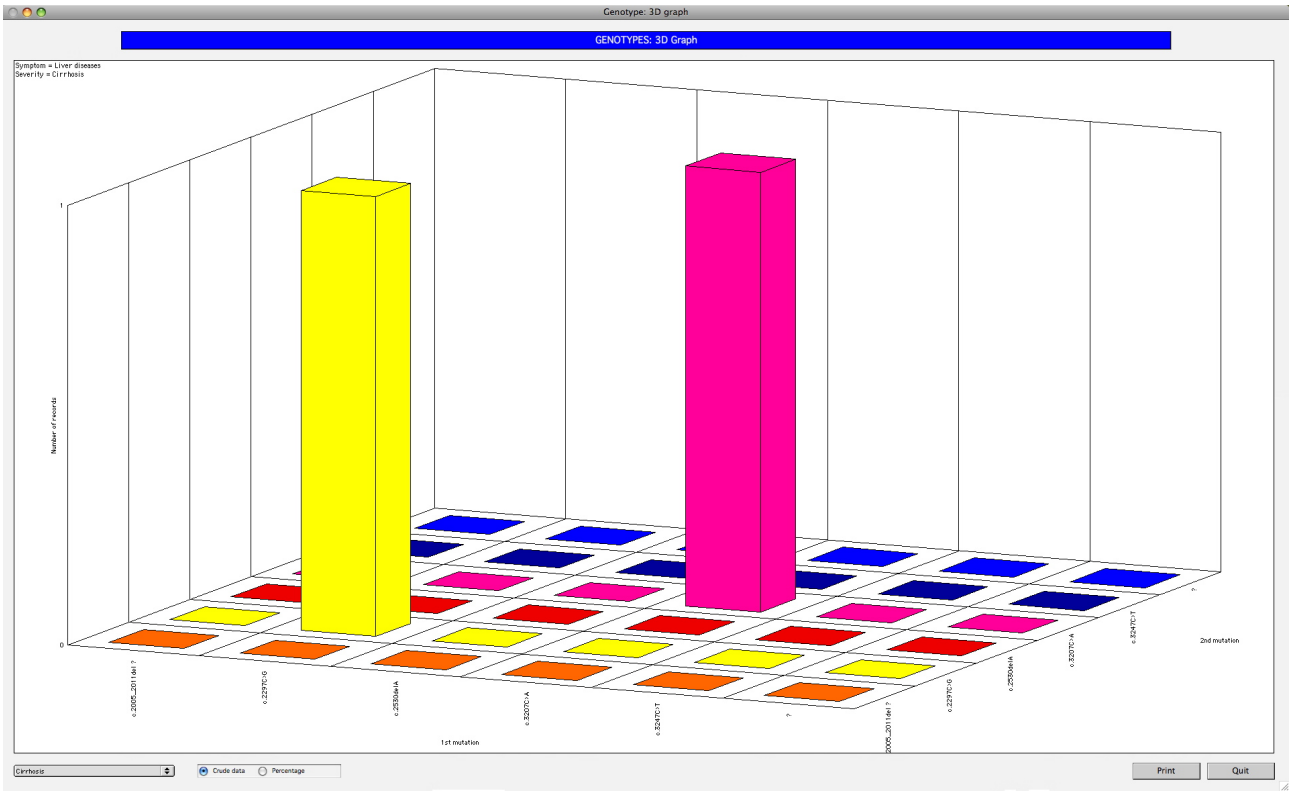
mutations.
For each patient, the to mutations are classified both at the class and at the severity levels in simple 2D tables. In the example presented here, 146 patients have 2 alleles harboring a class II mutation. A click on this box displays the list of corresponding genotypes on the large table on the right.

3) 3D Graph

As for the previous function, the "3D Graph" function can only be used for recessive mutations associated with clinical data.
The user can first select a subset of records and then a specific symptom.
The UMD® software will then select all patients with data for the selected symptom. It will then extract the various severities reported for this symptom and then display a 3D Graph with the various genotypes for allele 1 (x-axis) and allele 2 (y-axis) and the number of patients with the corresponding severity (z-axis).
The user can use the "left", "right", "up" and "down" arrows in order to modify the 3D-graph display.



A pop-up menu on the left bottom corner gives access to the other severities associated with the symptom.



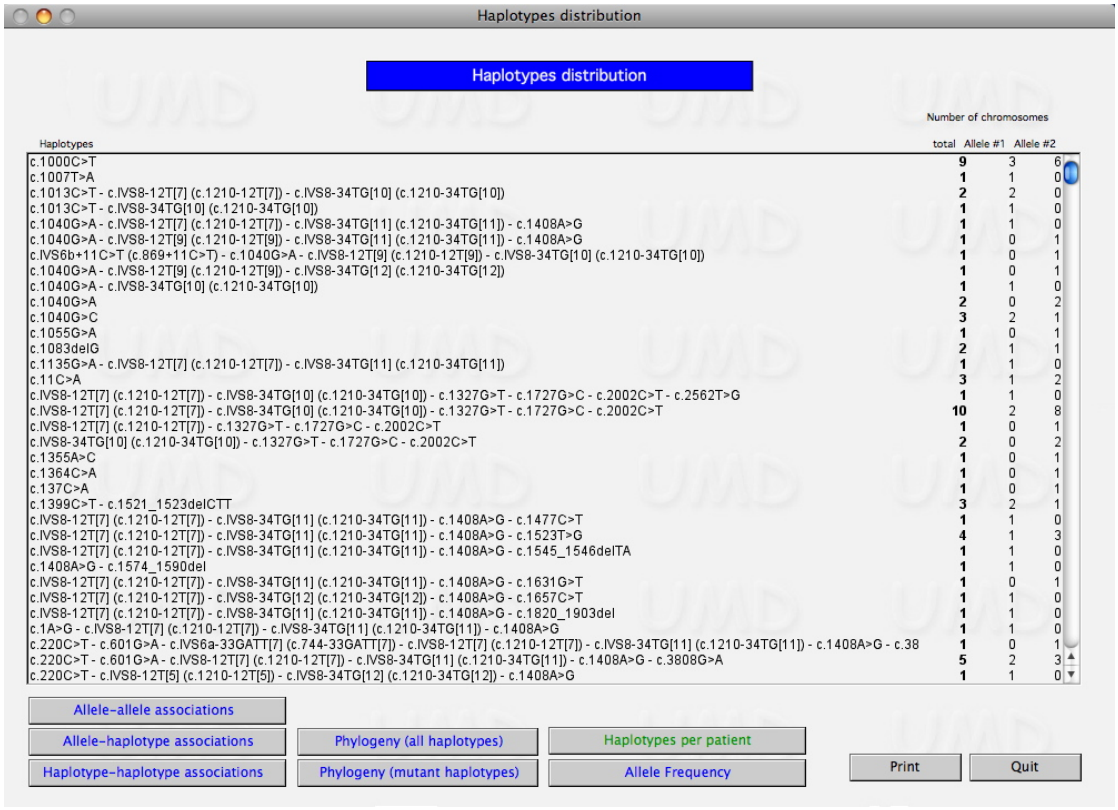
A second parameter (crude data / percentage) allows to modify the display z-axis format. In this example, only 2 homozygous ATP7B genotypes are associated with a cirrhosis.

VIII-i- The "Haplotype" menu

This menu is only accessible for curators who use a specific format that includes the "Allele" field. In addition all variations should be reported (mutations and polymorphisms) with the allele annotation.

1) Haplotypes distribution

The user can select a specific group of records using the usual search interfaces. The UMD® software will then reconstruct the haplotypes for each patient when data about alleles are

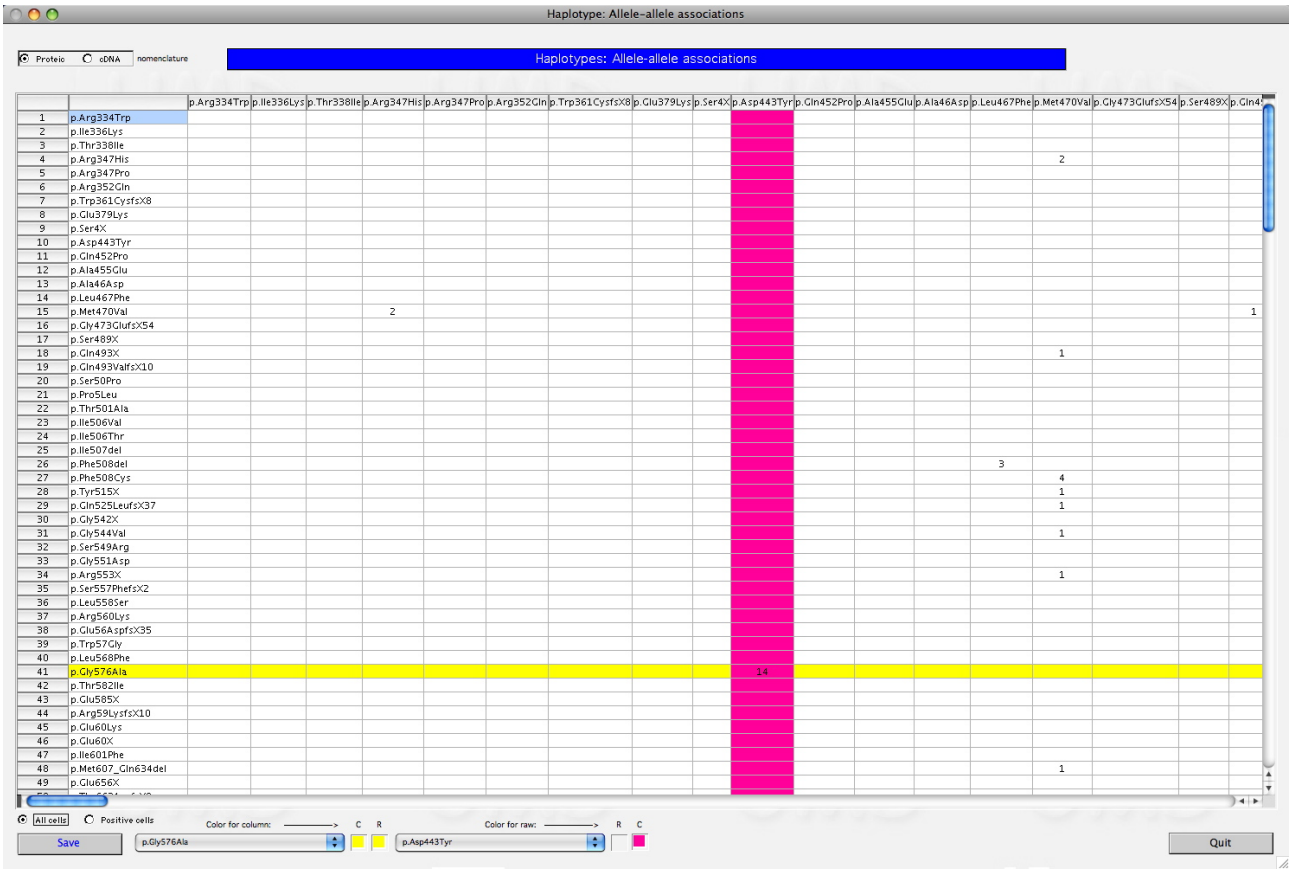


available. The first screen lists all haplotypes and the number of occurrences.
Seven buttons are available at the bottom of the screen:

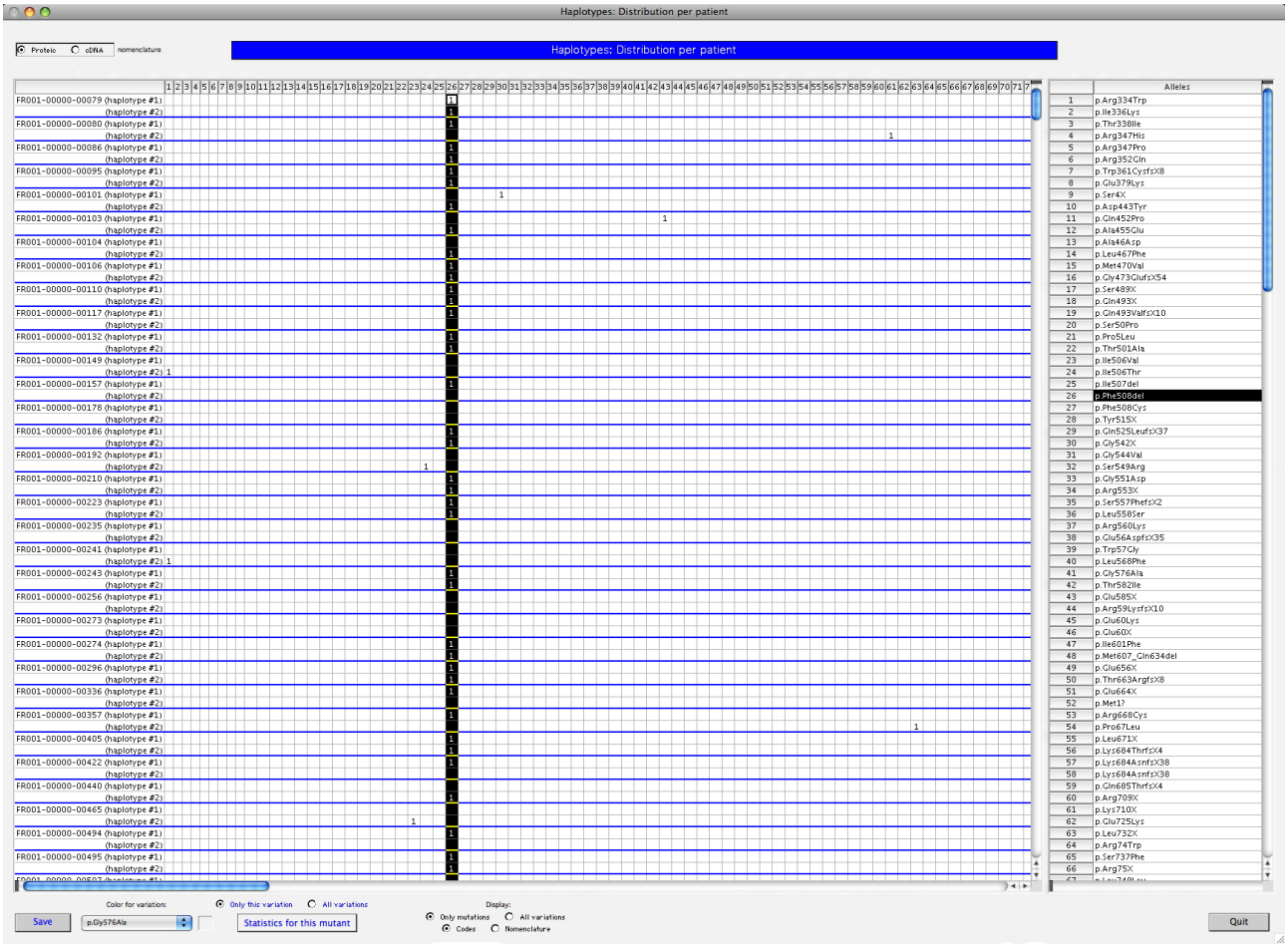
- The "*Allele-allele associations*" button is used to build a table allowing to cross alleles. Various parameters can be used to color rows and columns corresponding to one or 2 alleles;
- The "*Allele-haplotypes associations*" button is used to build a table allowing to cross alleles with haplotypes. Various parameters can be used to color rows and columns corresponding to one or 2 alleles or only positive cells;
- The "*Haplotype-haplotype associations*" button is used to build a table allowing to cross alleles with haplotypes. Various parameters can be used to color rows and columns corresponding to one or 2 alleles or only positive cells;
- The "*Haplotypes per patient*" button is used to build a table allowing to cross patients haplotypes with alleles. The user can underline a specific variation with a color from the 256 pallet. He can also choose to display only mutations or all variations. In

addition the c. or p. international nomenclature can be used or only a code to simplify the table. Finally, the "Statistics for this mutant" button allows to get details about this variation: number of homozygous and heterozygous and associated variations in cis or trans.

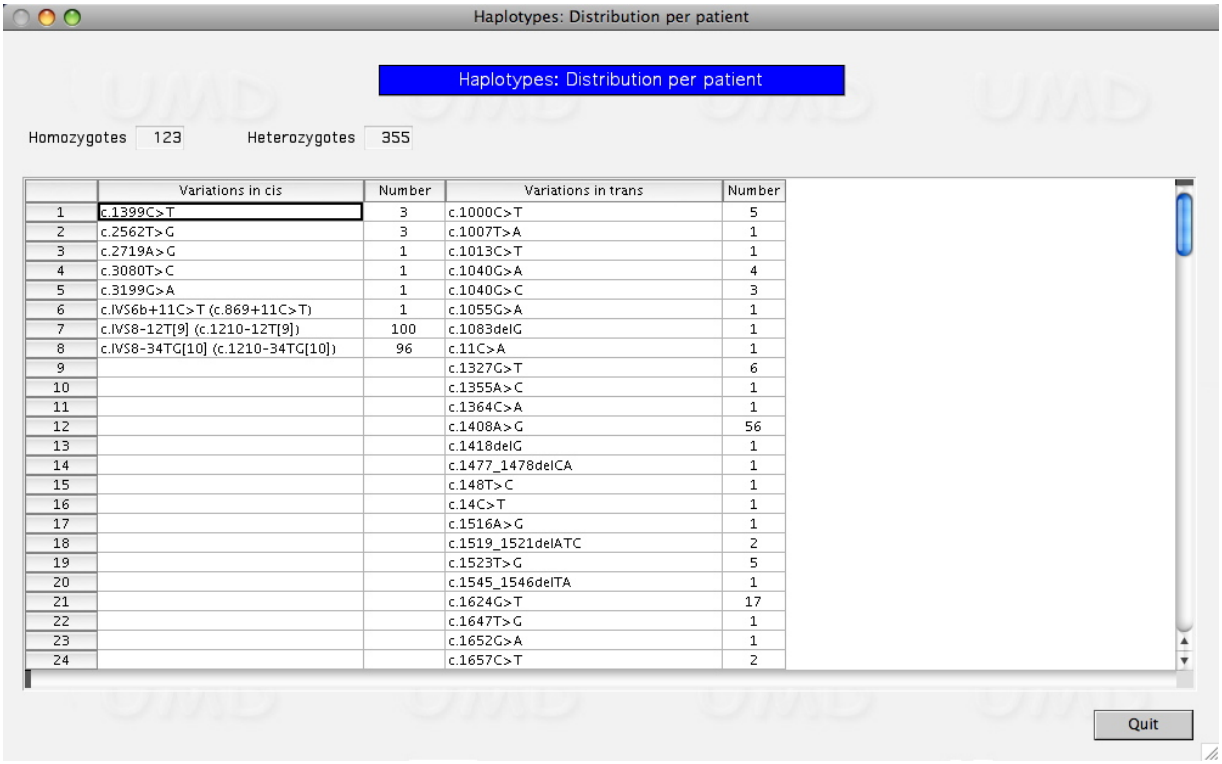
- The "*Phylogeny (all haplotypes)*" button is used to reconstruct the phylogenetic tree of the various chromosomes observed within patients. You need to install the dnamlk software prior to use this function. The UMD® software first prepare compatible files and launch the dnamlk application. An alert message ask you to indicate when the computation via the dnamlk is completed in order to import the results into UMD and display a phylogenetic tree with color codes within UMD. ***Only available upon request***.
- The "*Phylogeny (mutant haplotypes)*" button is used to reconstruct the phylogenetic tree of the various chromosomes observed within patients. Only chromosomes harboring mutations are taken into account for this analysis. ***Only available upon request.***
- The "*Allele Frequency*" button gives access to a simple table listing all variations and their frequency in the group of records selected by the user. The c. or p. nomenclature can be used for this analysis.



Example of the "*Allele-allele associations*" function. Two mutations have been color-coded.

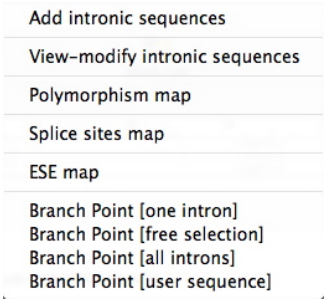


Example of the "*Haplotypes per patient*" function. The column corresponding the p.Phe508del mutation has been selected. Most patients are either heterozygous or homozygous for this mutation. Note that only the mutations are displayed. A click on the "*Statistics for this mutant*" button give the following results:



	Variations in cis	Number	Variations in trans	Number
1	c.1399C>T	3	c.1000C>T	5
2	c.2562T>G	3	c.1007T>A	1
3	c.2719A>G	1	c.1013C>T	1
4	c.3080T>C	1	c.1040G>A	4
5	c.3199G>A	1	c.1040G>C	3
6	c.IVS6b+11C>T (c.869+11C>T)	1	c.1055G>A	1
7	c.IVS8-12T[9] (c.1210-12T[9])	100	c.1083delG	1
8	c.IVS8-34TG[10] (c.1210-34TG[10])	96	c.11C>A	1
9			c.1327G>T	6
10			c.1355A>C	1
11			c.1364C>A	1
12			c.1408A>G	56
13			c.1418delG	1
14			c.1477_1478delCA	1
15			c.148T>C	1
16			c.14C>T	1
17			c.1516A>G	1
18			c.1519_1521delATC	2
19			c.1523T>G	5
20			c.1545_1546delTA	1
21			c.1624G>T	17
22			c.1647T>G	1
23			c.1652G>A	1
24			c.1657C>T	2

VIII-j- The "Introns" menu

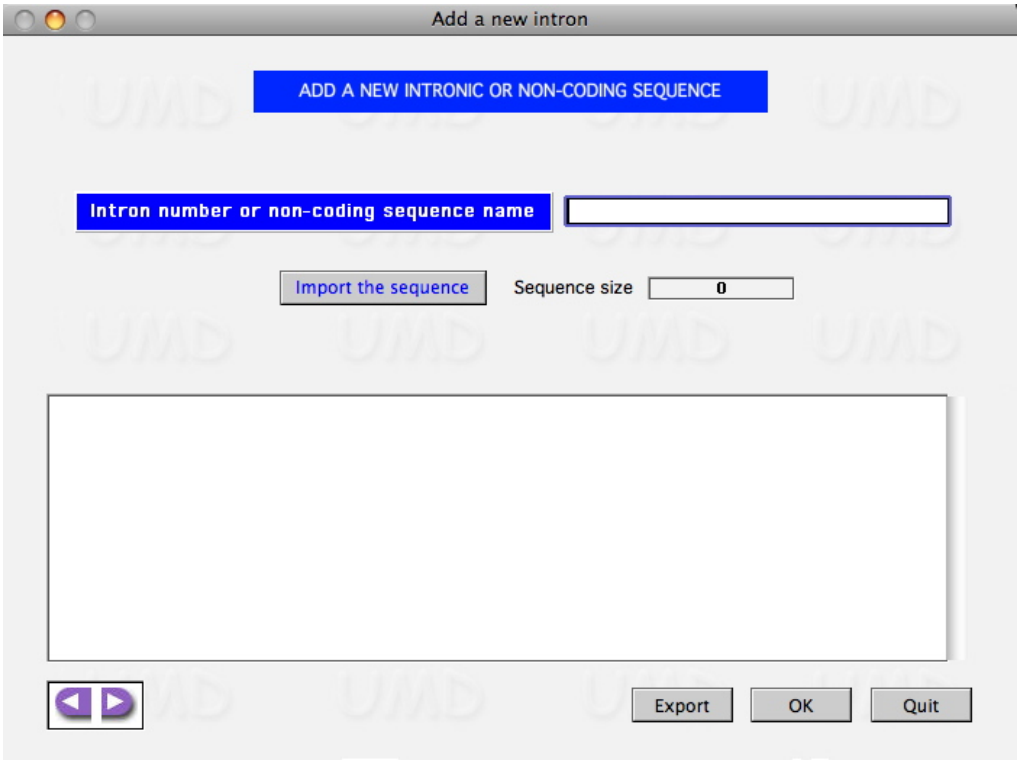


This menu contains 9 functions related to non coding sequences (introns and 5' and 3' regions).

1) Add intronic sequences

This function is used to add manually non-coding sequences into the database. If your gene of interest is available via NCBI, UCSC or Ensembl, we recommend that you import data using the "*Import polymorphic markers*" function from the "*File*" menu.

Important: All introns should be labeled "*Intron #xx*" where xx stands for the intron number. Other regions should be labeled "*3' region*" and "*5' region*". If you use other names, these sequences will not be recognized by the software.

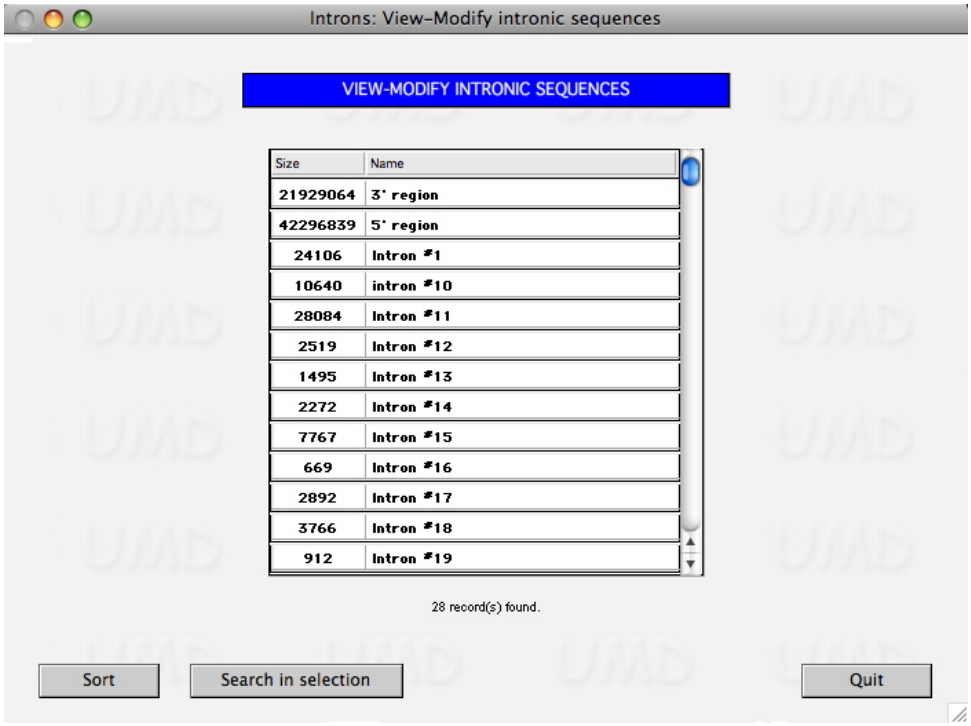


The "Import the sequence" button allows the import of a text sequence with only A, T, G or C characters (no numbers or end of line symbols). After the import the sequence size is displayed for control and the imported sequence can be visualized. Because only 32000 characters can be displayed at a time, 2 arrows allow to navigate within the sequence.

2) View-modify intronic sequences

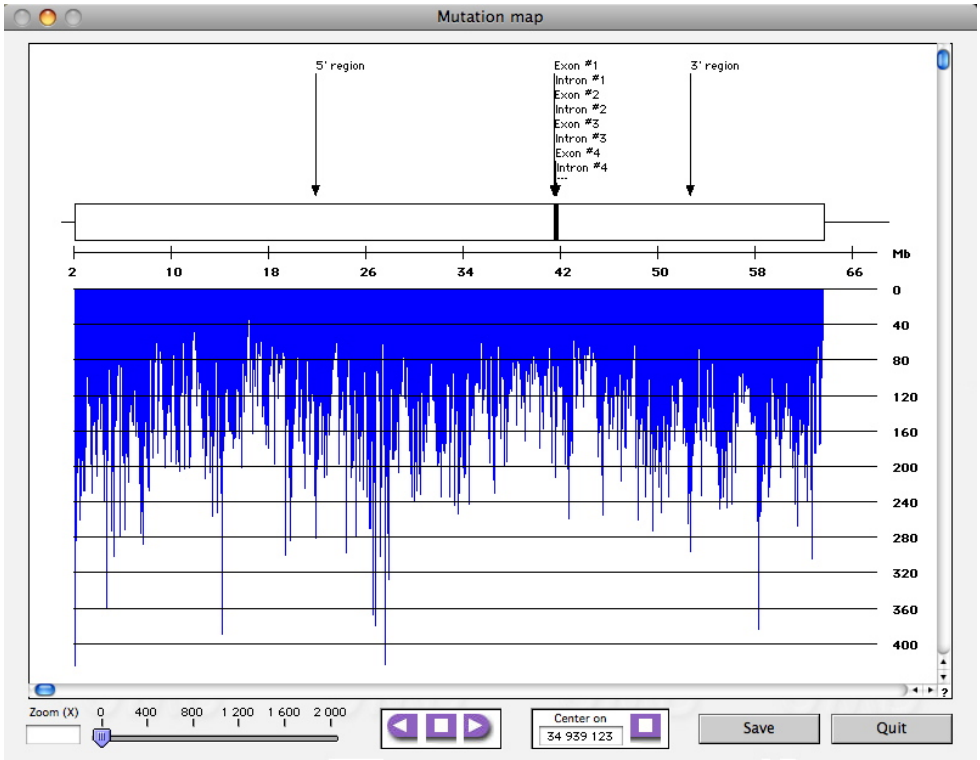
This function is used to display non-coding reference sequences used by the UMD® software and eventually modify these sequences.

A simple list of all records is presented. A click on the chosen non-coding sequence will open a screen similar to the one used for creation of intronic sequences (see previous function).



3) Polymorphism map

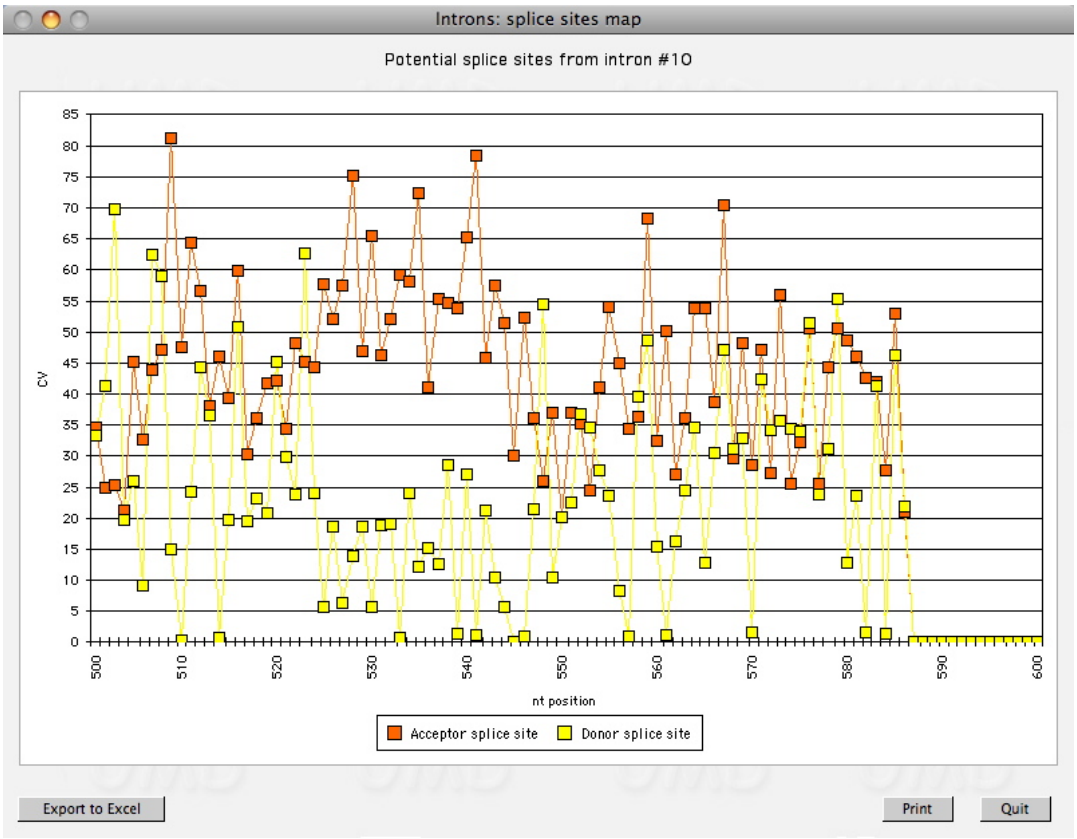
This function can be used only if polymorphic markers have been imported through the "*Import polymorphic markers*" function from the "*File*" menu. A graphical display of these markers is displayed:



A zoom option allow to get more details on a specific region of the genome. A zoom of x2000 will allow to click on the blue lines and display the corresponding information about SNPs and ESTs.

4) Splice sites map

This function is used to search for potential acceptor and donor splice sites within a non-coding sequence. It is usually used when an intronic mutation is predicted to activate a cryptic splice site. The user can then search for a complementary splice site in order to predict the potential cryptic exon that could be inserted in the cDNA. Remember that potentially active splice sites have aCV >70 (for more information, refer to the Human



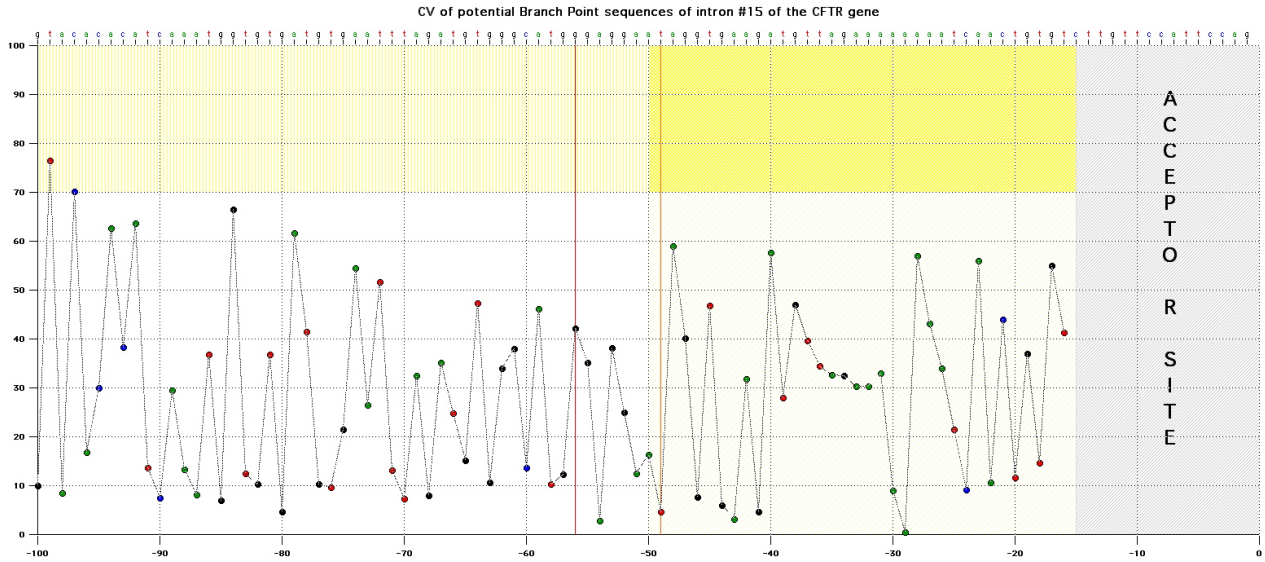
Splicing Finder tool Nucleic Acid Research, 2009, April).

5) ESE map

This function is used to search for potential auxiliary splicing signals using the ESE-Finder matrices. For a more sophisticated analysis, we recommend that you use the HSF tool.

6) Branch point [one intron]

This function is used to search for the branch point of an intron. The user selects one intron and the UMD® software will search for branch point (for more information, refer to the Human Splicing Finder tool Nucleic Acid Research, 2009, April).



Branch points are localized most frequently in the yellow region (-50 to -15), but can also be localized in the light yellow region (-100 to -51). Note that active branch point signals should have a value >70. The red and orange vertical lines correspond to the exclusion zone from Gooding et al.

7) Branch point [free selection]

This function is similar to the "*Branch point [one intron]*" function except for the sequence definition that can be freely chosen by the user. This is particularly useful in rare situations where the branch point in localized upstream to the 100 bp region or when looking for branch point of cryptic exons.

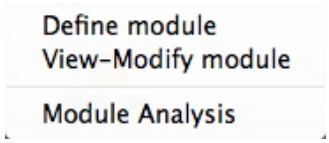
8) Branch point [all introns]

This function is similar to the "*Branch point [one intron]*" function except that all introns are computed simultaneously. This is usually used to get a global view of potential branch point localization for a single gene.

9) Branch point [user sequence]

This function is similar to the "*Branch point [one intron]*" function except that the user can search for branch point from any sequence of 100 bp.

VIII-k- The "Modules" menu



This menu contains 3 functions related to the definition and analysis of repeated modules within a single gene. It requires that structural domains have been defined in order to be used.

1) Define module

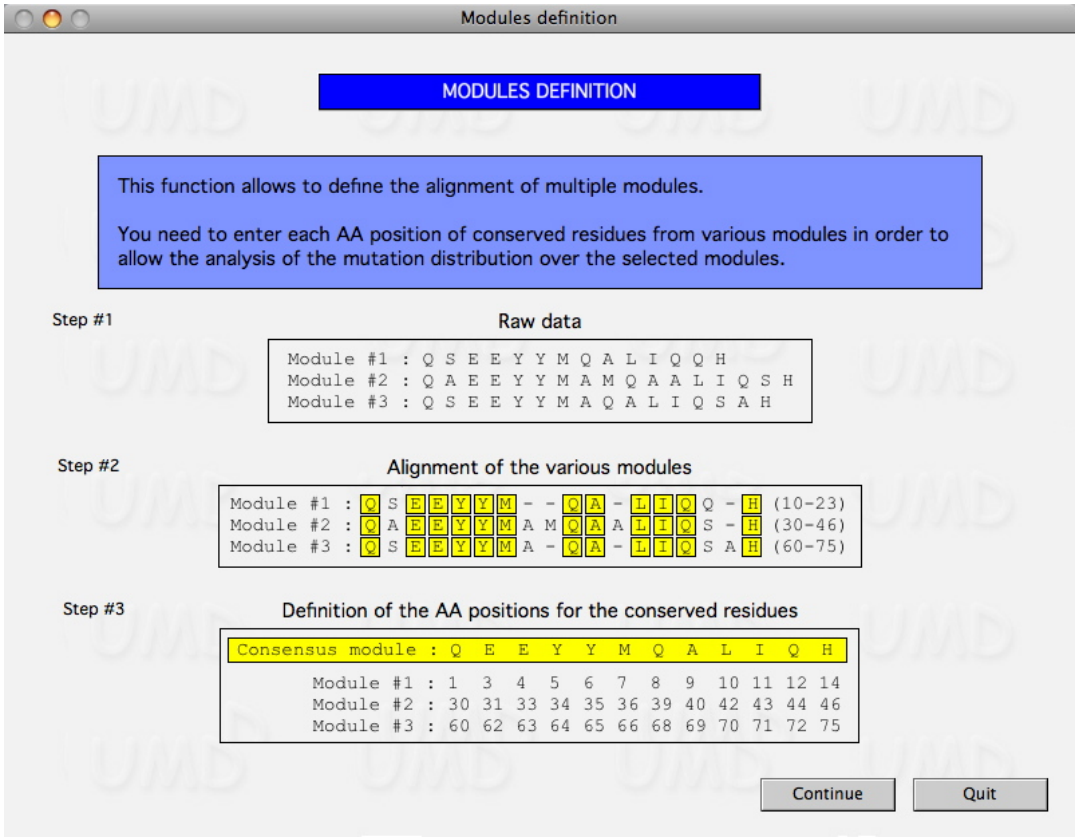
This function allows the alignment of multiple modules from the same protein. The user needs to convert the raw data into a consensus module.

The first step is to perform the alignment of the various modules using one of the numerous available tools.

The conserved and variant positions are then defined (step#2).

In the third step, the user will indicate to the software to position of each residue from a module. For example the module #3 has the "Q" conserved residue at position 1, a variant residue at position 2, the "E, E, Y, Y, M" conserved residues at positions 3 to 7, a variant residue at position 8...

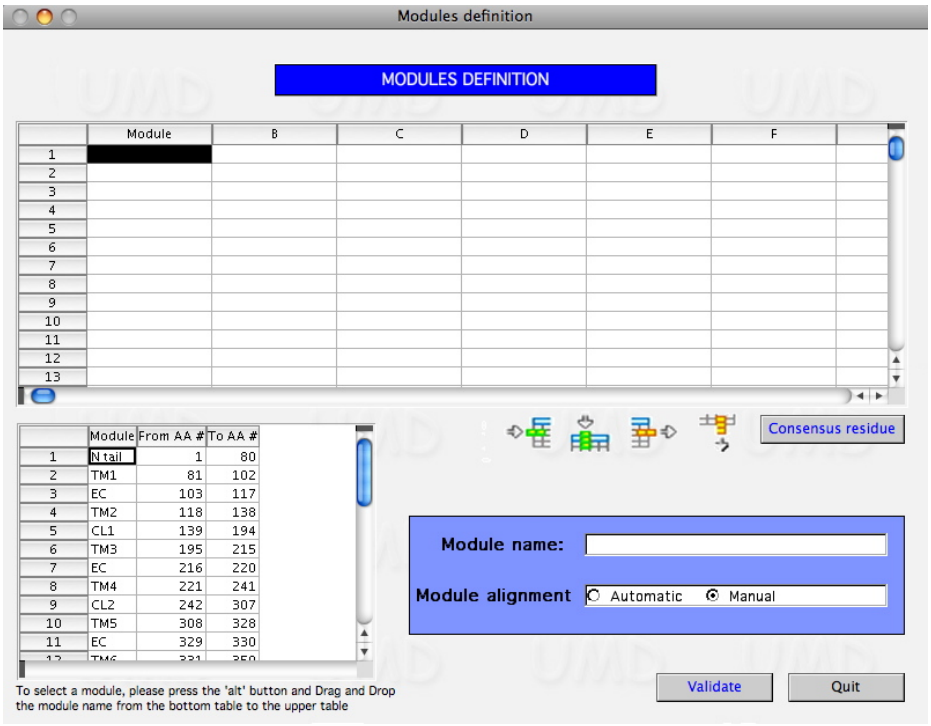
Note that position 8 (variant residue) can correspond to a single residue (module #3) or 2



residues (module #2) or event be absent (module #1).

In the second screen, two tables are available to create the consensus module. The bottom table contains all available structural domains. They can be selected and drop into the upper table (Alt+mouse). Each column should contain either a conserved residue or a * symbol to indicate that it is a non-conserved region.

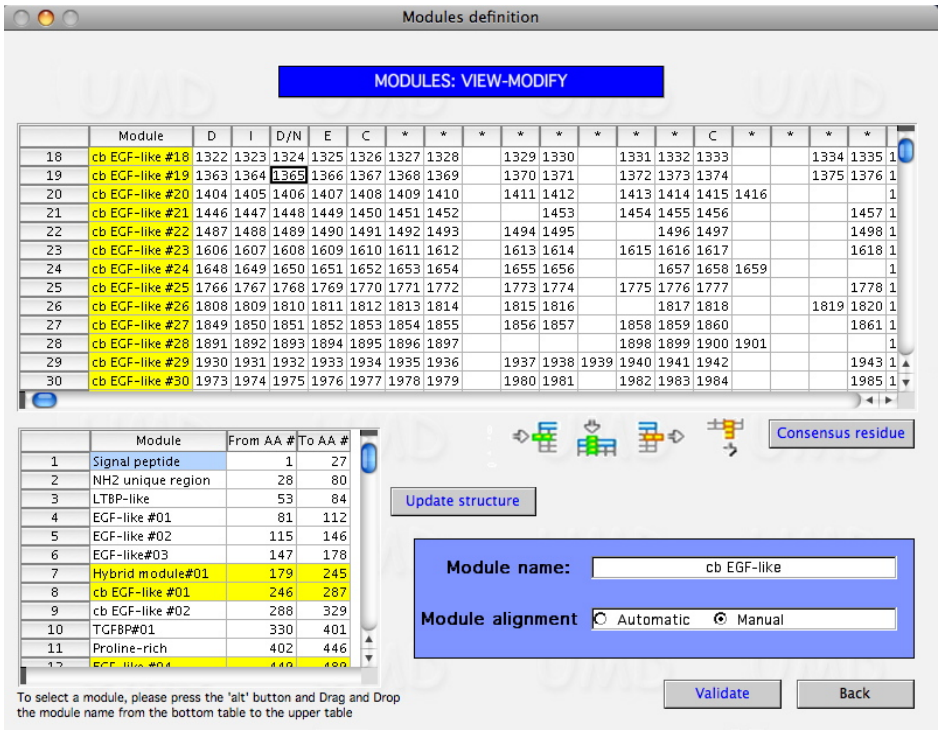
Use the "*Consensus residue*" button to annotate each column. Enter the position of the residue within the selected module in the corresponding cell (see the example from the "*Vie-Modify modules*" function).



Each module alignment should have a unique name "Module name".

2) View-modify module

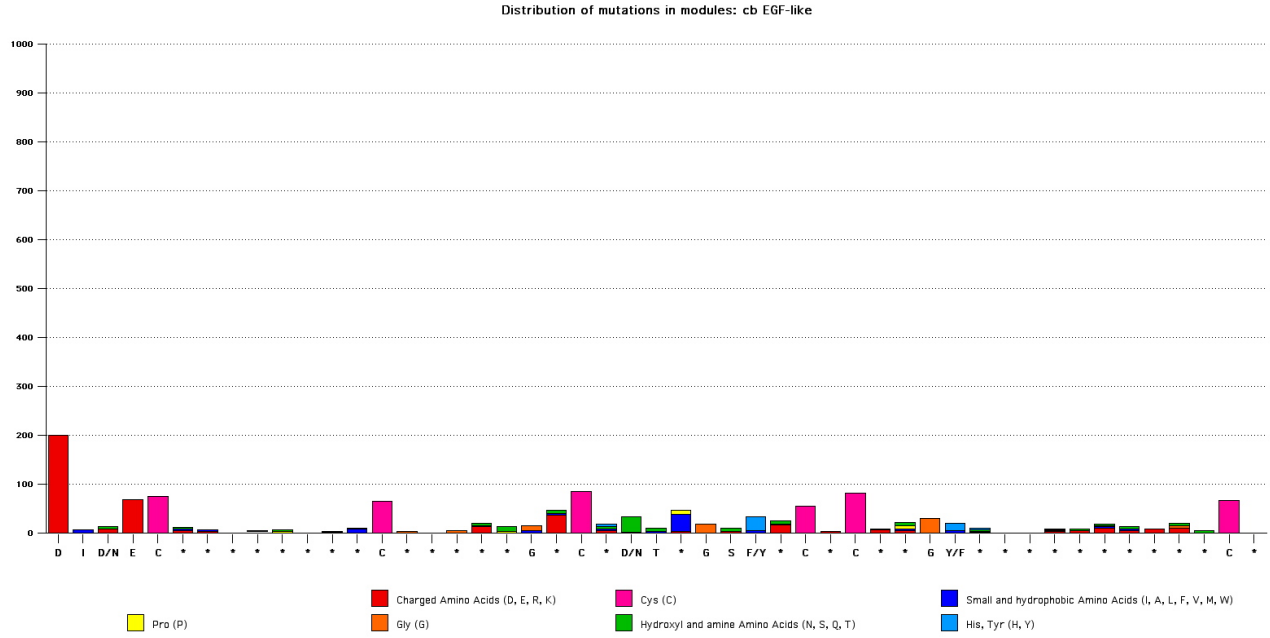
This function allows the modification of created modules alignments.



3) Module analysis

This function is used to study the distribution of mutations localized in different modules that have been aligned in a "*Module*" (see previous functions).

The user can select a subset of records prior to launch this analysis.



Amino acid with similar properties are grouped in one of the 7 groups and are displayed with a specific color code. The number of mutations affecting either a conserved residue or a variant regions are combined from all modules and displayed in the consensus module.

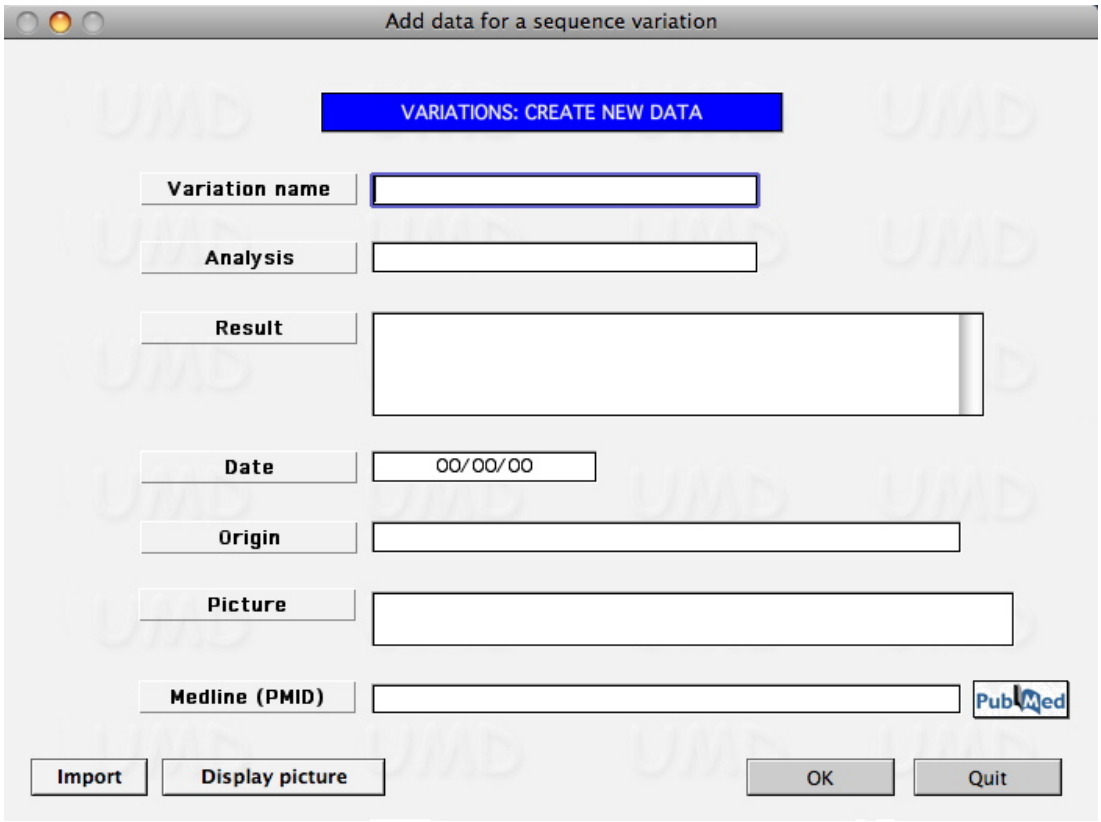
Note that in this example, mutations affect conserved Cys (C) residues and the first Asp (D) residue.

VIII-I- The "Variation" menu

This menu contains 4 functions related to the annotation of variations.

1) Add data for a variation

This option is used to collect evidences associated to the effect of a specific variation. Multiple evidences can then be collected for a single variation.
The variation should be described with the c. or p. international nomenclature systems in order to keep a coherence with data generated in the "*Mutation*" table.
The method of analysis can be stored in the "*Analysis*" field and the corresponding results in the "*Result*" text field.
In addition the date and the origin of the data can also be collected. If these data have been published, the PubMed ID can be store in the "*Medline (PMID)*" field and a direct link to the abstract is available via the "*PubMed*" button.
Finally a picture can be associated to this record using the "*Import*" and the "*Display picture*"



buttons.

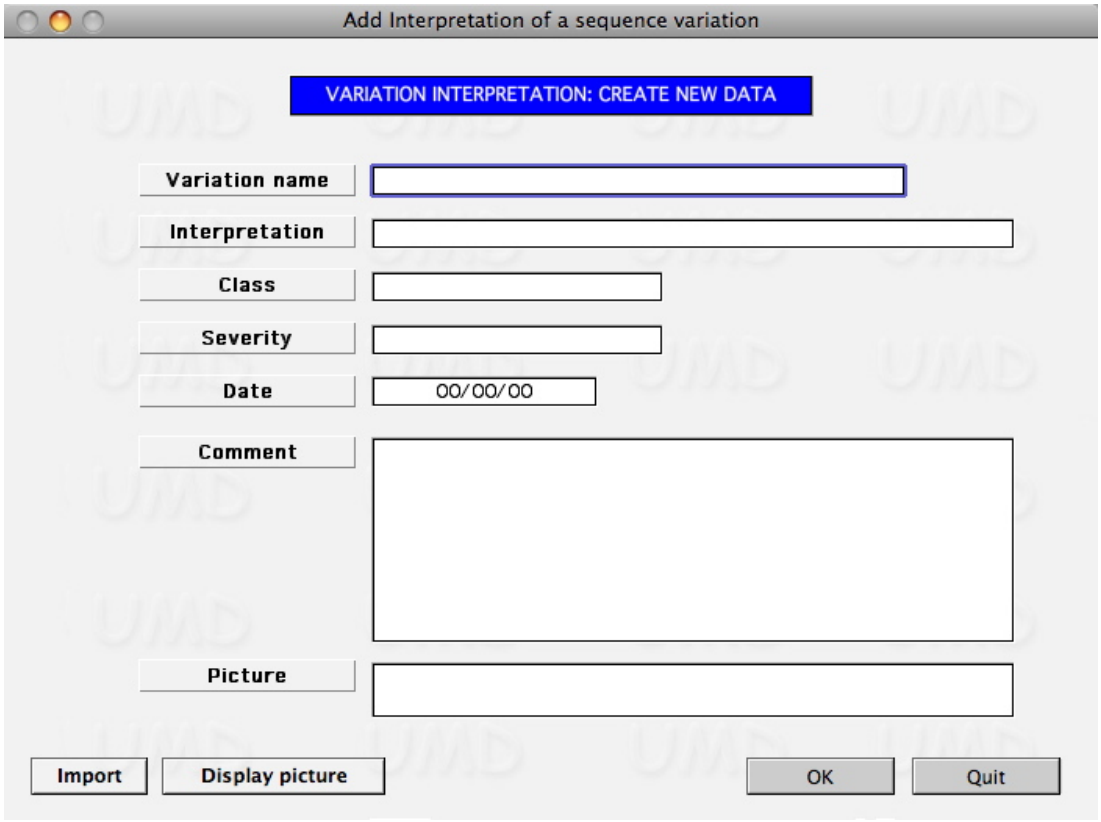
2) Variation data list

This option is used to view or modify the Variation records.

3) Add variation interpretation

While the previous "Variation data" table philosophy is to collect all evidences related to a specific variation, the philosophy of the "Variation interpretation" table is to conclude about the status of this variant. Therefore only one record per variant should be included in this table.
This table includes the following fields:

- The "*Variation name*" should be described with the c. or p. international nomenclature systems in order to keep a coherence with data generated in the "*Mutation*" table;
- The "*Interpretation*" should be done using the international consensus: "*Probably pathogenous*", "*Probably non pathogenous*", "*Strong evidence for being pathogenous*", "*Strong evidence for being non pathogenous*" and "*Unclassified variant*";
- The "*Class*" can be used when such system has been defined for a specific protein;
- The "*Severity*" can be used when such system has been defined for a specific protein in relation to a specific disease;
- The "*Date*" is used to store the date when the interpretation (consensus) was performed (reached);
- The "*Comment*" text field is used to describe specific features about this variation;



- The "*Picture*" field is used to store the picture file pathway. The "Import" and "Display picture" buttons are used to download and visualized this picture.

4) Variation interpretations list

This option is used to view or modify the Variation interpretations.

IX- CONCLUSION

The UMD® software was designed by Dr. Christophe Béroud more than 17 years ago in order to fill the gap between core databases such as GenBank and the day-to-day benchwork leading to discoveries of mutations responsible for human genetic diseases. Thierry Soussi and Gwenaëlle Collod-Béroud played a major role in the evolution of this system since these early days and the creation of the first UMD-LSDBs.

UMD is now maintained and developed at the INSERM UMR827 directed by Pr. Mireille Claustres. A bioinformatics team directed by Christophe Béroud is regularly developing new tools, Dr. Dalil Hamroun playing a major role in this process.

This whole adventure has been made possible by the numerous collaborators and users who have chosen to use our tool and develop UMD-LSDBs. Because they have new questions and problems related to their genes, we are frequently confronted to research challenges that can benefit from large scale analysis and bioinformatics.

The emergence of new genotype-based therapies has led to the involvement of clinicians and the creation of new functions to increase or understanding of the natural history of diseases, the search for genotype-phenotype associations and the possibility to select group of patients for clinical trials.

Because the field of Locus Specific DataBases and Knowledgebases (to which belongs the UMD system) is rapidly growing, we are facing new challenges to better understand the relationships between mutations and human diseases and ultimately move towards the predictive medicine.

Finally, we would like to give special thanks to:

The **AFM** (**A**ssociation **F**rançaise contre les **M**yopathies) that has supported this work other the years;

The European Community Sixth Framework Program FP6 under grant agreement number 036825 - the **TREAT-NMD** Network of Excellence;

The European Community Seventh Framework Program FP7/2007-2013 under grant agreement number 200754- the **GEN2PHEN** project.