Retrieval of variants-related articles from PubMed

Written by: Nathalie Lachenal
Directors: Anne-Lise Veuthey, Lina Yip and Violaine Pillet
University of Geneva
2006/2007
Summary

Single amino acid polymorphism (SAP) is the type of mutation most related to human diseases. The Swiss-Prot knowledgebase records 29’684 variants in 5’222 proteins (Release 51.4, 9 Jan. 2007). At least one reference is provided for each polymorphism. This information is however not up-to-date. Indeed, the increasing number of publications on variants and their linked pathologies renders the update time-consuming and difficult.

We proposed here an automatic method to retrieve variants-related articles that might be useful to update Swiss-Prot information. For information retrieval, we first determined a set of four Perl regular expressions representative of the different SAPs notations by reviewing the existing references in Swiss-Prot. This set of regular expression was further validated using a medical annotation corpus. For each protein of interest, we then queried PubMed using protein names, gene names and a set of mutation-related keywords to find related abstracts. The retrieved documents were parsed and their variants content was extracted using the four regular expressions. The position of the extracted variants was further validated using Swiss-Prot sequence information. For the valid documents, we classified the extracted polymorphisms as known variants or potential new variants.

The trial of the retrieval process on the existing references of Swiss-Prot showed a specificity of 84%, and a minimal recall of 48%. From the complete set of human variant proteins in Swiss-Prot, we extracted and validated 23’520 potential new references. These data are accessible on the corresponding Swiss-Prot Variants pages through the link additional references.
# Table of contents

## Summary

1. Introduction
   1.1. Biological meaning of single amino acid polymorphisms
   1.2. Medical annotation in Swiss-Prot
   1.3. Text-mining techniques in the frame of references retrieval

2. Methods
   2.1. Corpora
      2.1.1. Swiss-Prot references corpus
      2.1.2. Medical annotation corpus
   2.2. Regular expressions and rules
   2.3. Variants retrieval program structure
   2.4. Positional validation of extracted variants
      2.4.1. Swiss-Prot information required for validation
         2.4.1.1. Positional validation with RP and FT lines
         2.4.1.2. Potential shifts of sequence calculation
         2.4.1.3. Positional validation with SQ lines
      2.4.2. Documents evaluation by positional validation
         2.4.2.1. Variants notations normalization
         2.4.2.2. Documents validation procedure
         2.4.2.3. Potential new variants evaluation
   2.5. Automatic procedure pipeline
   2.6. Evaluation metrics

3. Results
   3.1. Extraction with regular expressions
      3.1.1. Results on the Swiss-Prot references corpus
      3.1.2. Results on the medical annotation corpus
   3.2. SAPs retrieval from PubMed queries
      3.2.1. Pattern matching analysis for the retrieved documents
      3.2.2. Positional validation outcome
      3.2.3. Information stored in the database

4. Discussion

5. Conclusion

Appendix

Bibliography
1. Introduction

The high throughput techniques are in constant evolution and provide a tremendous amount of biological data. Scientific publications remain the primary source for communicating these results. It is difficult to retrieve information and to keep the biological databases content up-to-date with this ever-growing volume of new articles. Literature on single amino acid polymorphisms (SAPs) is no exception, despite the fact that information related to the molecular and phenotypic consequences of SAPs is essential for the scientific community to understand the nature of these polymorphisms and their possible relationship to human diseases. In Swiss-Prot knowledgebase, while all SAPs have at least one corresponding PubMed reference, this information is not up-to-date. In this project, we propose an automatic method based on text-mining to retrieve relevant articles on variants from PubMed in order to facilitate the update of information on SAPs.

1.1. Biological meaning of single amino acid polymorphisms

The understanding of single nucleotide polymorphisms (SNPs) begins at the genomic level. The term SNP is used to indicate a variety of nucleotide substitutions at a certain locus; the most common nucleotidic polymorphisms are base insertion, deletion, or biallelic polymorphism, not frameshifts or nonsense mutations. SNPs that are involved in multi-allelic modifications are very rare and are generally associated with a specific genetic disease. The frequency of polymorphisms in exons is approximately 1 in every 1’250 base pairs in the human genome [Liu and Regnier, 2003]. When a polymorphism occurs in a gene coding for a protein, it can cause a SAP, it means the substitution of amino acids in the sequence of the protein. The fact that a SNP is in a coding gene does not always leads to a SAP, because alterations in the second and third bases of a codon may not induce a change of amino acid due to the degeneracy of the genetic code. Thus, SNP frequency will be higher than SAP frequency. Less than 1% of all SNPs result in a variation in proteins. Nevertheless, the SAP is the type of mutation most related to human diseases. One of the best-known examples is sickle cell anemia, a disease characterized by a single glutamate to valine mutation at position 6 in the β-globin chain. This single mutation causes the polymerization of hemoglobin and as a consequence this SAP leads to premature death of the red cells and results in anemia [Ferrer-Costa and al., 2002].

Because of the physiological importance of SAPs, numerous studies have been published on the subject. Single amino acid polymorphism data are increasing with the use of high throughput techniques, for instance restriction fragment length polymorphism mapping (or RFLP mapping) that allows the discovery of several polymorphisms of one single protein [Pearson, 2006]. Thus, there are numerous publications on SAPs in which information on novel linked-diseases, phenotypes, or recently discovered molecular implications of the variation can be retrieved.

1.2. Medical annotation in Swiss-Prot

Swiss-Prot is a knowledgebase of protein sequences which provides, for each protein entry, a unique accession number (AC), bibliographical references, taxonomic data, comments and sequence features. The annotation in Swiss-Prot is a manual process consisting in description of
protein function, post-translational modification, protein domains and sites, variants, structure, etc. [Bairoch and al., 2005]. Protein annotation is a complex task, as it requires considerable human time and expertise. One major process in annotation is literature mining to find out the interesting information.

The Swiss-Prot knowledgebase records 29’684 variants in 5’222 human proteins (Release 51.4, 9 Jan. 2007). Polymorphisms are especially important for the research community. Consequently, SAPs are treated by a specific medical annotation process in Swiss-Prot. Diseases and polymorphisms are described in specific fields. Comment lines (CC), prefixed by the term disease or polymorphism, summarize medical and biochemical knowledge on the disease or the polymorphism. For example in the human androgen-receptor (AC P10275):

```
CC   -!- POLYMORPHISM: The poly-Gln region of AR is highly polymorphic and
CC   the number of Gln varies in the population (from 17 to 26). A
CC   smaller size of the poly-Gln region may be associated with the
CC   development of prostate cancer.
```

Besides, the variants of each protein are listed in the Swiss-Prot features lines (FT) of the entry, in the order of their position in the amino acid sequence. For example, a mutation in the Tau protein (AC P10636) is described as:

```
FT   VARIANT     285    285       D -> N (risk factor for progressive
FT                                supranuclear palsy).
FT                                /FTId=VAR_010340.
```

In these FT lines, we have the position of the change 285, the description of the polymorphism D -> N, some information on the etiology of the variant, as well as the FTId, the unique identifier of the polymorphism.

A summary of all available information on such reported variants can be found in the Swiss-Prot Variant pages that can be accessed via a hyperlink provided through FTIds. The database providing additional information on each Swiss-Prot SAP is ModSNP [Yip and al., 2004]. Apart from general information such as the amino acid change, the position, the effect of the variant and its association to disease, each ModSNP entry provides new structural information on the variant. In particular, when available, there are 3D-models generated by homology modelling on which we can visualize the mutation directly on the protein structure.

Each Variant page is composed of five sections:
1. general information describing the protein,
2. information on the variant,
3. structural information on the variant,
4. references for the polymorphism,
5. cross-references for the variant to other specialized databases such as OMIM (Online Mendelian Inheritance in Man, on NCBI), dbSNP (Database of single nucleotide polymorphism, on NCBI) or HGVbase (Human Genome Variation database).

Moreover, each Swiss-Prot entry provides relevant literature citations. The Swiss-Prot references are high-quality documents selected by the annotators and providing interesting complementary information on the protein.

The following example shows how an article on polymorphism is cited in the HLA class II histocompatibility antigen entry (AC P01903):

```
RN   [10]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA], AND VARIANT LEU-242.
RC   TISSUE=Blood, and Colon;
RX   PubMed=15489334; DOI=10.1101/gr.2596504;
```
The article entitled “The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC).” mentions the variant Leu-242 of the human HLA class II histocompatibility antigen. Through the PubMed unique identifier of the article (PMId), we can access the abstract stored in MEDLINE, the repository of citations and abstracts of biomedical research articles. A specific line, the RP line, describes the contents of the cited reference of an entry. These lines provide a short description of the essential information of the article, for instance: nucleotide or protein sequence, post translational modifications, protein structure, function etc. In particular, if the citation corresponds to an article on polymorphism, it contains the term “variant” as shown in the example. By using these existing references on polymorphisms in Swiss-Prot, we will get acquainted with the way authors notate SAPs and determine an automatic method to be used to efficiently retrieve variants-related articles from PubMed.

1.3. Text-mining techniques in the frame of references retrieval

Because there is a need in the scientific community to find out and extract information, such as SAPs, text-mining approaches have been developed. In the mid-1980s the first manual text-mining techniques appeared, but the field had progressed quickly during the past decade with the technological advances. Text-mining is interdisciplinary; it combines computational techniques with linguistics and statistics [Krallinger and al., 2005]. Biological or medical text-mining is a bioinformatic approach of the field that allows computer to gather, store and analyse biological or medical data [Dobrokhotov, 2005].

The science of searching information in texts or searching for the documents themselves is called information retrieval (IR). Another branch of text-mining that will be used here is information extraction (IE); it means the scanning of written documents and the population of a database with the extracted information [Spasic and al., 2005].

The text documents analyzed by text-mining are modelled according to computational and statistical processes. The most advance techniques of the topic are based on a mixture of natural language processing (NLP) and machine learning algorithms. These techniques combine artificial intelligence (AI) and linguistics. Machine learning permits the development of inductive or deductive learning algorithms. Thanks to this learning capacity, NLP allows an automated understanding of human languages. The combination of AI techniques trained on representative text datasets allows the computer to create rules and patterns to “understand” and “learn” text structure.

Syntax analysis by a computer is called parsing, it means the analysis of a sequence of tokens, which are categorized blocks of text, in order to determine grammatical structure of the text document with respect to a given formal grammar. As language structure and rules on writing are defined, information can be retrieved from texts very efficiently [Erhardt and al., 2006].

The aim of this work is to develop an automatic method to retrieve the most relevant information on variants in articles from PubMed. This method is inspired by text-mining techniques using rule-based linguistic methods. We are going to manually determine specific lexical fields, rules, or notation conventions that will allow the computer to recognize the articles on SAPs in the literature and to assign each variant a set of related articles.
2. Methods

2.1. Corpora

2.1.1. Swiss-Prot references corpus

Existing references on variants in Swiss-Prot were used as starting corpus. As these references had already been selected according to their relevance, they were used as reference points (Gold standard) for the elaboration and training of regular expressions. They were also used to compare the performances between different approaches.

The PubMed identifiers (PMIds) of the Swiss-Prot references mentioning the term “variant” in at least one of their RP line were obtained by parsing Swiss-Prot entries using Swiss-knife, the available Perl package to parse Swiss-Prot [Fleischman and al., 2001]. The list of PMIds was then used to query the scientific literature database PubMed to retrieve the titles and abstracts of the corresponding articles. As the same article might be cited in many Swiss-Prot entries, some redundancy would be found in this set of retrieved references. Publications that had been already present in the dataset were not kept. However, for every repeated document, the list of all AC and their corresponding RP lines were recorded.

The retrieved data were reformatted so that they would be easily readable by the program that was to parse the documents and extract positional information on SAPs. We chose to put the title and the abstract together on a single line bounded by tags: a start tag <tag> and an end tag </tag>. The reformatted line became:

⇒ <RT>title</RT><AB>abstract</AB>.

To every title and abstract lines, we associated the AC numbers, the PMID, the corresponding RP lines and the other information on the article such as journal and authors. The documents were isolated by a separator character // followed by the end of line character.

2.1.2. Medical annotation corpus

This dataset contained 2’188 classified documents for 32 proteins. It was used to train a tool that assisted information retrieval in Swiss-Prot medical annotation process [Dobrokhotov, 2005]. The documents were retrieved from PubMed with a query containing keywords (mutation, variant, and polymorphism) associated to the gene and protein names and their synonyms. A group of medical annotators was asked to read the title and abstract of each article, and assign each document to a category: relevant or irrelevant, according to their usual reviewing criteria when annotating a protein entry. In the cases where it was not possible to confidently assign the document to any of the two categories, it was classified as unclear.

This corpus was used as a complementary corpus to train, refine and improve the regular expressions. It contained, in addition to relevant articles, a set of irrelevant documents that would enable us to test the discriminating character of the regular expressions and adapt them to the cases where abstracts were poor in information.
The documents had been classified as:

<table>
<thead>
<tr>
<th></th>
<th>Good</th>
<th>Bad</th>
<th>Unclear</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of articles</td>
<td>306</td>
<td>1'528</td>
<td>354</td>
<td>2′188</td>
</tr>
</tbody>
</table>

**Tab.1.** Classification of documents in medical-annotation corpus.

For the *Good* articles, we should note that some of them had already been parsed because they had been reported as references in Swiss-Prot and owned a RP variant line. These documents were still kept in order to maintain the integrity of the corpus.

The corpus contained more *Bad* documents than *Good* ones, but it reflected the case of a real annotation process. When we search articles in PubMed to obtain new information on Swiss-Prot variants, we will meet such unbalanced situations.

### 2.2. Regular expressions and rules

Perl regular expressions for variant information extraction were defined by manual review of approximately hundred abstracts from the Swiss-Prot references corpus. The following patterns correspond to the different observed notations mentioning the position of single amino acid polymorphisms:

<table>
<thead>
<tr>
<th>Pattern n°1</th>
<th>Expression style</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern n°1</td>
<td>1 letter code amino acid – number – 1 letter code amino acid</td>
<td>▪ <em>L45H</em>,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▪ L-45-H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▪ L(45)H</td>
</tr>
</tbody>
</table>

| Pattern n°2                          | 3 letters code amino acid – number – 3 letters code amino acid or nucleotide triplet | ▪ Leu 45 -> His,            |
|                                      |                                                                                   | ▪ Leu(45)His,               |
|                                      |                                                                                   | ▪ CTT(45)CAT                |

| Pattern n°3                          | number – 3 letters code amino acid – 3 letters code amino acid or nucleotide triplet | ▪ 45Leu -> His,             |
|                                      |                                                                                   | ▪ 45CTT>CAT,                |
|                                      |                                                                                   | ▪ 45Leu - - > His           |

| Pattern n°4                          | Sentences                                                                            | The leucine 45 to histidine polymorphism. |

**Tab.2.** Overview of the four regular expressions.

The pattern n°1 is the most frequent pattern that corresponds to the standard for notation of variants according to mutation nomenclature [den Dunnen and al., 2001]. Other notations are accepted, such as the pattern n°2, that possess the same structure of information as pattern n°1,
but the amino acids are expressed in three letters code or maybe as nucleotidic codons. Pattern n°2 and n°3 differ in the arrangement of the number and the amino acids.

In many titles and abstracts of articles on SAPs, the polymorphism is not expressed by such expressions, but is explained in sentences. That is the reason for which we defined pattern n°4 that retrieved sentences containing variants. By searching the occurrence of a certain number of keywords in a sentence, we can be reasonably confident that it mentions a SAP. We consider that a sentence matches the pattern, if it contains all these features:
- a number (position of the variant),
- a term describing the variation, like: variant, mutation, or polymorphism.
- an amino acid in full-letters, 3 letters code or codon notation with a term describing the position in the amino acid chain, like: place, position, amino acid, residue, etc. Or if the sentence contains no term on position but mentions two different amino acids.

For example the sentence, *We can observe a change of alanine to valine at position 345 due to [...]*. This sentence is considered as possessing information on variants. (exhaustive vocabulary list, in Appendix 1.)

The patterns n°1, n°2 and n°3 depict clearly the direction of the amino acid change. With pattern n°4 it is more difficult to determine the direction of the variation (whether it is from X to Y, or from Y to X) or which number in the sentence describes the position in the protein sequence. We defined a rule to read these sentences based on the order of its terms regarding some vocabularies on polymorphisms. For instance, the sentence: *We observe a change of the glycine to an alanine at position 45*. Here, the rule allows recognizing that the first amino acid cited is the one that is replaced and the second is the variant, because they are separated by the preposition to. There is the word position that indicates that the number following this term is the position of the change. In this example the syntax is clear for the program and it recognizes the mutation: G45A. (Vocabulary terms listed in Appendix 2.)

But sometimes, part of the information is missing or the sentences are more complicated. In these cases, we used a more random way to find out the variant in the sentence. We retrieved every amino acid in the sentence and every number, and formed all possible combinations. Among these random associations of characters, we tried to find which polymorphism was really expressed in the sentence. The necessary differentiation step will be explained later (section 2.4.).

Additional rules on pattern n°1 were also determined. Indeed, we have observed that some gene or protein names adopt this pattern (e.g. A2M). To reduce the number of this type of errors, the information retrieved with the pattern n°1 is compared to the name of the protein of interest or of the corresponding gene to remove these potential false positives.

Some other common errors were associated with pattern n°1. For example, the expressions G2M and G1S which matched the first variant notation, but if they were present in an article talking about cell cycle, these expressions corresponded to mitotic phases transitions and not to SAPs. Thus, every time we obtained G2M or G1S with the pattern n°1, we verified if terms such as phases, cell cycle, etc. were found in the abstract. If it was the case, the expressions G2M and G1S were removed from the match. Other expressions that were systematically removed according to the context were E2F for epidermal growth factor, the M24R filter model or the cell lines C3H, T47D or C13T. The complete list of expressions and contexts is in Appendix 3.
2.3. Variants retrieval program structure

The Perl program for SAPs-related information retrieval took into account the previously defined regular expressions and rules. The program consisted of many filtering steps that eliminated progressively articles that did not contain the indispensable information. This structure permitted one to avoid testing the regular expressions on irrelevant text extracts.

Steps in the SAPs retrieval program:

1. **Input:** The program read the title and abstract of every document as a single line.

2. **First filter:**
   If the title and the abstract did not contain any number, the article was considered not containing positional information on variants and was discarded. The notation of a sequence position with a number expressed in full-letters has not been observed, so we do not consider this case.

3. **Test of the three first regular expressions:**
   In every article, each expression corresponding to one of the patterns (n°1, 2 or 3) was tested as follows:
   
   I. First, we kept only the units of text containing one or three letters, associated to a number and to another block of one or three letters.
      \[ \Rightarrow aa1 - number - aa2 \text{ or } number - aa1 - aa2 \]
   
   II. Then, there was a check to verify that what we called \( aa1 \) was an amino acid:
       - in one letter code for pattern n°1,
       - in three letters code for patterns n°2 and 3,
       - in codons for patterns n°2 and 3:
          \[ \Rightarrow \text{As we did not want to recover nonsense mutations that truncated the amino acid chain, every time there was a codon, we verified that it was not a stop codon (TAG, TAA or TGA). If it was a stop, we did not test the following step, because the expression was not describing a SAP.} \]
       Otherwise, if the \( aa1 \) did not correspond to an amino acid, the expression was not tested further and was eliminated from the set of publications.
   
   III. The second check verified if the \( aa2 \) was also an amino acid in one or three letters code or a codon; but this residue must be different from the \( aa1 \), or else it meant that the pattern described a silent mutation.

If the regular expressions confirmed that an extract of text corresponded to one of the SAPs notations, we classified the article as potentially describing the found polymorphism.

4. **Second filter and test of pattern n°4:**
   If the title and the abstract did not contain any vocabulary terms describing polymorphisms, such as: variants, polymorphism, mutation, substitution, change, etc. The program would not test the last regular expression which needed a text containing these words.
   For pattern n°4 the unit title and abstract was cut into sentences with a tokeniser [Durant, 2003]. This part of the research program retrieved sentences describing polymorphisms using a certain
number of keywords. The sentences considered to possess information on variants were returned by the retrieval program.

5. Output:
At the end, each article was classified as containing or not SAPs. The program provided, for every informative article, a list of SAPs or of sentences containing variants in case of pattern n°4. An abstract might mention several notations forms of the same or many different polymorphisms.

2.4. Positional validation of extracted variants

2.4.1. Swiss-Prot information required for validation

The list of patterns extracted above for each article contained either known variants, potentially new variants and false positives. This step used Swiss-Prot-based information to differentiate between the known and the unknown variants, as well as between the true and false positives in order to confidently assign articles to the right SAP.

2.4.1.1. Positional validation with RP and FT lines

We considered two different cases:
1. SAPs retrieval on the Swiss-Prot references corpus to test the performances of the variants retrieval program,
2. SAPs retrieval in the framework of the automatic method aiming to update variant information in ModSNP.

In the first case, we compared, for each document, the variants extracted from text to the ones listed in the corresponding RP lines in order to assess if the expressions and sentences extracted from these abstracts were reliable polymorphisms. Indeed, most RP lines mentioned the changing amino acid and the position, e.g. RP VARIANT STHE LEU-299. (AC P23415), except reviews on variants, e.g. RP REVIEW ON VARIANTS. (AC P05067). In such cases, we used the protein entry’s FT VARIANT lines and compare their content to the extracted SAPs of the document. This comparison allowed one to know if we retrieved relevant variants information using our retrieval process.

In the second case, FT VARIANT lines were used to determine if an article was a potential new reference for an already identified variant. If one of the SAPs extracted from an abstract corresponded to an already known polymorphism in Swiss-Prot, we assigned the document to the category of potential new references for this variant.

2.4.1.2. Potential shifts of sequence calculation

If correspondence between SAPs extracted from the documents and the RP lines or the FT VARIANT lines was not found, 3 explanations are possible:
- There is a sequence numbering problem causing position mismatch,
- These SAPs are new variants,
- The expressions extracted from text are errors, they do not represent SAPs.
Assuming that the mismatch between the variants reported in RP or FT and the ones extracted from the abstract was caused by a different sequence numbering, we took into account some features that could possibly explain these differences.

First, the number of the position of the initiator methionine was, not long ago, considered to be the position n°0 in Swiss-Prot. Now, the first methionine is position n°1 in the amino acid chain. Depending on the numbering chosen by the authors of the articles, we accepted an error in the position of +/- 1 amino acid.

Second, some potential shifts of sequence were considered as the sequence in Swiss-Prot corresponds to the unprocessed chain. Several interesting characteristics of sequence that could be removed or added to the amino acid chain are reported in FT lines. These features can help us to calculate shifts, for example loss of signal or transit sequence in the author’s protein.

e.g. **FT TRANSIT 1 25 Mitochondrion (By similarity).**
This FT line shows that there is a transit sequence 25 amino acids long that allows the transfer of the protein in the mitochondrion. This element of the protein chain can be removed or added and consequently modifies the numbering of positions. We can calculate the position as if such shifts occurred in the author’s sequence given the length of the modification found in the FT TRANSIT and SIGNAL lines.

There are also the FT VAR_SEQ features. The term VAR_SEQ describes the sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation or ribosomal frameshifting. The VAR_SEQ can be of two different types:

- A sequence is modified:
  
  **FT VAR_SEQ 653 672 VATSNPGKCLSFTNSTFTFT -> ALVSHHCPVEAVRAVHP**
  
  **FT**
  
  **FT**
  
  **FT**
  
  **/FTId=VSP_003786.**

- A part of the sequence is missing:
  
  **FT VAR_SEQ 673 913 Missing (in isoform 2).**
  
  **FT**
  
  **/FTId=VSP_003787.**

(AC Q8R2H7)

These features can also serve for calculating shifts of position. An isoform can cumulate many VAR_SEQ. Therefore, all the features concerning the same isoform of a protein were taken into account for shift calculation.

Adding such shifts of sequence permitted recognizing variants that had not been numbered the same way as in Swiss-Prot.

**2.4.1.3. Positional validation with SQ lines**

If, after comparing with all known variants of the protein by taking into accounts possible differences in the position numbering, it still remains extracted SAPs expressions that do not correspond to the already reported variants of the protein, we could suppose that they are new polymorphisms.

To decide if these variants expressions are new SAPs or errors, we used the amino acid sequence of the protein. This information is found in Swiss-Prot SQ lines which provide the sequence corresponding to the unprocessed precursor.

Each SAPs pattern specifies which amino acid at which position had been replaced in the variant. Thus, we verified if we found this amino acid at the indicated position in the sequence, by taking into account the potential shifts of sequence. The variants having a correspondence with the sequence, but that are not stored in Swiss-Prot, are regarded as potential new variants for the protein of interest.
2.4.2. Documents evaluation by positional validation

2.4.2.1. Variants notations normalization

The main process of the validation procedure was to compare the variants from the documents to those reported in Swiss-Prot. The polymorphisms notations extracted from the articles were first normalized for comparison.

The normalized form was chosen to be: 1 letter amino acid code, followed by the position of the variation, then the one letter code amino acid after the change. e.g. $G45A$, that expressed the conversion of a glycine at position 45 to an alanine. The specific notations of pattern 1 ($G-45-A$), pattern 2 ($Gly45 \rightarrow Ala$), pattern 3 ($45GLY\rightarrow ALA$), could easily be converted to the standard form ($G45A$).

For pattern n°4, the cases were treated as explained in chapter 2.2, by generating either a single variant notation or all possible combinations of every mentioned amino acid and number. The random character of this last extraction of information would be taken in account in the evaluation step.

Concerning the Swiss-Prot information on variant, we convert:

a) The RP lines, for testing the Swiss-Prot references corpus. RP VARIANT STLHE ALA-45 becomes 45A. As the RP lines mention just the polymorphic amino acid, we compared only this information.

b) The FT lines, FT VARIANT 45 45 G $\rightarrow$ A. is easily converted to $G45A$.

2.4.2.2. Documents validation procedure

Document validation for Swiss-Prot references corpus:

The validation process for the Swiss-Prot references corpus was special, in the sense that these articles had already been read by curators and the reported SAPs were referenced in the corresponding RP line. To assess the relevance of the SAPs retrieval using the regular expressions, we used the following validation process based on RP lines for each variant extracted from a document:

![Validation process for Swiss-Prot references](image)

**Fig. 1.** Validation of the variants extracted from Swiss-Prot references. At the step 4, the variants were compared again to RP and FT VARIANT lines before reaching step 5. The documents were classified according to the degree of verification of their extracted SAPs.

The documents in which all SAPs content were valid, were considered to be well analysed by the SAPs retrieval program. The abstracts in which a part or the whole SAPs content remained unverified were regarded as unsuccessfully scanned by the regular expressions. Indeed, given the
fact that the existing references in Swiss-Prot were manually curated, there was a high probability that the variants information that could not be verified with the RP and FT lines was irrelevant. Knowing the number of well-retrieved documents from the Swiss-Prot references would enable us to estimate the performances of the SAPs retrieval approach.

Document validation for new references retrieval:

The validation process for new references was slightly different than that for the Swiss-Prot corpus. One of the main differences was that we could no longer verify the extracted information with RP line. We therefore verified the SAPS content with the FT lines in order to check if the newly retrieved references reported an existent variant in Swiss-Prot. If this was not case, the SAPs content was further checked with the SQ line to determine if it would be a new variant. The validation procedure was outlined below:

Fig. 2. Validation of the variants extracted from an automatically retrieved document. At step 3, if the extracted SAPs did not correspond to a known variant even after shift calculation, the expressions was compared to the sequence with or without shifts before reaching step 5. The documents were classified according to their variants content.

At the end of the validation process, the remaining documents were all linked to at least one known or unknown SAP of the protein of interest. This positional validation process was used at the end of the automatic method (section 2.5).

2.4.2.3. Potential new variants evaluation

For the documents classified as mentioning potential new SAPs, we associated a “confidence” score to every potential new variant, according to its pattern. We attributed less credit to the pattern n°4 because of the random character of its determination.

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Error tolerance</th>
<th>Confidence level</th>
</tr>
</thead>
<tbody>
<tr>
<td>n°1, 2 or 3</td>
<td>+/- 1 amino acid</td>
<td>high</td>
</tr>
<tr>
<td>n°4</td>
<td>+/- 1 amino acid</td>
<td>medium</td>
</tr>
<tr>
<td>n°1, 2 or 3 shift</td>
<td>+/- 1 amino acid</td>
<td>low</td>
</tr>
<tr>
<td>n°4</td>
<td>shift +/- 1 amino acid</td>
<td>very low</td>
</tr>
</tbody>
</table>

Tab.3. Confidence level of the potential new variants.

In every category, high, medium, low or very low, the potential new variants were classified according to the number of articles in which they were mentioned. The possible new SAPs were ordered according to this classification:
The potential new variants for each protein will be displayed in this order. This classification allows having an idea of the value of the information presented.

2.5. Automatic procedure pipeline

The starting point of the automatic retrieval was a list of Swiss-Prot accession numbers of proteins of interest. We used the list of Human entries with polymorphisms or mutations in Swiss-Prot (Humpvar document, release 51.4, 9 Jan. 2007, http://www.expasy.org/cgi-bin/lists?humpvar.txt). For every protein, we searched new references on its SAPs using the method described below. All the programs performing the steps of the automatic retrieval had been written in Perl.

New references retrieval process:

For each AC number, we carried out all the following steps:

1. **Query on Swiss-Prot:**
   With the AC of the protein of interest, we sought in the corresponding Swiss-Prot entry and extracted information that would be used in the PubMed query and during the validation process:
   - The DE (for DEscription) and GN (for Gene Name) lines, containing the protein name, the gene name, and their synonyms.
   - The FT VARIANT, TRANSIT, SIGNAL and VAR_SEQ lines.
   - The SQ lines.
   - The list of PMIDs of the protein’s references with RP lines containing the term “variant”.

2. **Query on PubMed:**
   The query was formulated by using the list of gene and protein names and their synonyms, as well as a list of keywords: mutation, polymorphism and variant.
   The eUtils tool was used to retrieve all available articles on variants for each protein of interest [Khovayko, 2006]. The retrieved documents were in Medline format, they were converted to the format readable by the research program. At this step, the articles that had already been used in Swiss-Prot were removed from the corpus by using the PMIDs from the Swiss-Prot entries.
3. Searching variants in abstracts:
With the four regular expressions, the SAPs retrieval program sought for polymorphisms notations in the titles and abstracts. The results were displayed as a list of SAPs for each variant-related document. The documents in which we did not find any variant were removed from the list of potential new references.

4. Validation of articles as possible references for a variant:
In this step, we sought to classify the SAPs found in the documents into either known or unknown variants, or false positive (section 2.4.2.3). The same document could be used as reference for many different polymorphisms of the same or of different proteins. Moreover, an article could be present in more than one category, if it mentioned an already known variant and a new variant for example.

5. Storage in database:
At the end of the retrieval process, the resulting data (new references on already stored SAPs or references on potential new variants of known proteins) were stored in tables of the ModSNP postGreSQL database [Yip and al., 2004]. The references for already known variants were associated with the FTId of the SAP. And new variants were associated with the confidence score calculated above. For details on the content of the tables, please refer to appendix 4.

Fig. 4. Automatic method for retrieval of new references on variants. The figure showed, for each program, the input data and the queried database in grey, the used Perl modules in green-blue, the achieved tasks in purple, and the resulting PostGreSQL tables in dark blue.
6. Results display:
The new references for known polymorphisms were rendered accessible to public on the corresponding Swiss-Prot Variant page through a hyperlink titled “additional references” in the references section. The web-interface has been developed in collaboration with the “Providing structural information on variants” project. The potential new variants were not displayed, but were stored in the same table. All references will need a manual assessment to determine if their content is relevant. To facilitate the analysis of this information by the user, we provide a context for every polymorphism. This allows a quick check of the pertinence of the references looking to the environment of the variant in text.

2.6. Evaluation metrics

To estimate the performances of our automatic retrieval process on Swiss-Prot references corpus and on medical-annotation corpus, we calculated the precision, and the recall, which are two performance metrics usually used in information retrieval and extraction.

The precision is the fraction of true positives on the set of all positives results; it indicates the proportion of retrieved document that is relevant:

\[
\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}
\]

The recall is the fraction of true positives on all true results; it represents the proportion of documents considered to be relevant on all relevant documents:

\[
\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}
\]

These statistics were used to evaluate the efficiency of the retrieval of variants-related abstracts using our automatic method.

3. Results

3.1. Extraction with regular expressions

3.1.1. Results on the Swiss-Prot references corpus

The Swiss-Prot corpus consisted 5’778 references related on the variants of 4’651 human proteins. This corpus was sorted for the presence of polymorphisms using the SAPs retrieval program (Methods 2.2.). We retrieved all occurrences of the patterns n°1 - 4 in the title and abstract of these 5’778 articles. It was found that 3’010 documents (52%) contained at least one of the four patterns and 2’768 (48%) contained none of these patterns.

Among the 2’768 documents containing no patterns, 211 abstracts belonged to the subset in which no positional information on variants was present in their titles and abstracts, because they did not contain any number. We also noted the presence of 627 abstracts not detectable by the pattern n°4 as they contained no terms describing polymorphisms.
Among the 3'010 documents in which we found polymorphisms information, the pattern n°1 was the most frequently observed (55%), followed by the patterns n°4 (23%) and n°2 (19%). The pattern n°3 (3%) appeared to be the less effective.

**Graph.1.** The contribution of each of the four patterns to all expressions extracted

We evaluated the performance of each pattern in retrieving the correct information on the variant specifically annotated in the Swiss-Prot entry. We proceeded by comparing the extracted polymorphisms only to the ones found in RP lines. The proportion of verified variants from each pattern with regard to the total number of distinct patterns in the same category was calculated. The results are shown in Graph 2.

**Graph.2.** Evaluation of the performance of different patterns using RP lines.

We observed that pattern n°1, which was the most abundant, performed the best, with 71% verified by information mentioned in the RP lines of the entry. The pattern n°4 gave a more moderate result, with only 52% of the extracted polymorphisms verified by the RP lines.
From the 3’010 variants-related documents, 3’861 different variants expressions were extracted, meaning more than one per document. Therefore, we compared these extracted variants not only to the corresponding RP information, but also to the FT VARIANT lines to verify their relevance (Method 2.4.2.2.). We counted the number of validated variants with regard to the total number of extracted expressions:

<table>
<thead>
<tr>
<th>Unverified</th>
<th>476 SAPs</th>
<th>12.3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with RP</td>
<td>2’630</td>
<td>SAPs 68.1%</td>
</tr>
<tr>
<td></td>
<td>Exact match with RP</td>
<td>2’228 SAPs</td>
</tr>
<tr>
<td></td>
<td>Mismatch in position of +/-1</td>
<td>195 SAPs</td>
</tr>
<tr>
<td></td>
<td>Shift of sequence calculated</td>
<td>207 SAPs</td>
</tr>
<tr>
<td></td>
<td>Verified with FT VARIANT</td>
<td>106 SAPs</td>
</tr>
<tr>
<td></td>
<td>Mismatch in position of +/-1</td>
<td>55 SAPs</td>
</tr>
<tr>
<td></td>
<td>Shift of sequence calculated</td>
<td>594 SAPs</td>
</tr>
<tr>
<td></td>
<td>Shift only</td>
<td>80 SAPs</td>
</tr>
<tr>
<td></td>
<td>Shift &amp; mismatch of +1</td>
<td>33 SAPs</td>
</tr>
<tr>
<td></td>
<td>Shift &amp; mismatch of -1</td>
<td>94 SAPs</td>
</tr>
<tr>
<td></td>
<td>Shift only</td>
<td>336 SAPs</td>
</tr>
<tr>
<td></td>
<td>Shift &amp; mismatch of +1</td>
<td>148 SAPs</td>
</tr>
<tr>
<td></td>
<td>Shift &amp; mismatch of -1</td>
<td>110 SAPs</td>
</tr>
</tbody>
</table>

**Tab.4.** Verification of all different patterns found in the Swiss-Prot references with the details of the comparison to RP and then to FT VARIANT information. Shifts and tolerance of a difference of +/-1 amino acid in the variant’s position rescued ~25% of the information. We have to note that these data had been calculated before the change of the numbering of the initiator methionine in Swiss-Prot.
Among the 3'010 documents with variant patterns, we verified the patterns in each document according to the whole validation procedure for this corpus (Methods 2.4.2.2.). We considered that a document was completely verified if all the variant patterns were valid. The result is shown in Table 5.

<table>
<thead>
<tr>
<th>Documents verified with RP only</th>
<th>Documents verified with RP and FT adding shifts</th>
<th>Documents not verified</th>
<th>Total of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>2'074 (69%)</td>
<td>462 (15.3%)</td>
<td>474 (15.7%)</td>
<td>3'010</td>
</tr>
</tbody>
</table>

Tab.5. Results of the validation of the documents from Swiss-Prot reference corpus.

By manual review of 120 of the 474 documents containing still unverified expressions, we observed that they all contained bad matching patterns taken for variants by the SAPs retrieval program.

For 103 abstracts of these documents, the sentence detected by pattern n°4 did not give the position of the polymorphisms, even if the number is near the “position” term or the amino acids. Sometimes it was the position in the nucleotide sequence or totally irrelevant information such as atomic mass of a compound. It would be difficult to further restrict the sentence content; as this would lead to the loss of some sentences containing the right information.

In the remaining 17 documents, pattern n°1 did not correspond to what we looked for. For example, the expression \( A \rightarrow T \) referred to a change in the DNA sequence and thus highlighted the confusion between SNP and SAP notations. The retrieval program interpreted this expression as the substitution of an alanine in threonine in the amino acid chain. But it described, in fact, an adenine to thymine change at the position 456 in the nucleotide sequence. This mutation led certainly to a SAP, otherwise the reference would not have been reported with a RP variant line in Swiss-Prot. However, we could not interpret the effect of such SNPs at the protein level, so we did not take in account these variants.

The precision of retrieval with the regular expressions on the Swiss-Prot corpus was calculated. We define as false positives the documents containing still unverified patterns.

\[
\text{Precision} = \frac{2'536}{(2'536 + 474)} = 84\%
\]

The precision was good as expected, as this was the corpus used to determine the regular expressions. We would need to evaluate the performance of this extraction method based on other data (see below). As we had not examined all rejected articles (2’768) to determine if they were all true negative, we could not calculate the exact recall. We could however estimate the minimum recall of this extraction method by assuming that all rejected documents were false negatives.

\[
\text{Minimal recall} = \frac{2'536}{(2'536 + 2'768)} = 48\%
\]

Thus we had at least 48% of relevant documents retrieved on all relevant documents of the corpus. However, the manual review of approximately 200 documents that had been rejected by the retrieval program showed no false negative, thus leading us to think that the exact recall would be quite good.
3.1.2. Results on the medical annotation corpus

The medical annotation corpus contained Bad documents that would not be used to annotate a protein entry. This bad information allowed us to test the discriminating character of the SAPs retrieval program. We sought for variants with the four regular expressions in the titles and abstracts of the 2'188 articles of this pre-classified corpus. We obtained 385 documents (17%) in which we found at least one variant notation and 1’803 documents (83%) in which there was no pattern. The documents were classified as follows:

<table>
<thead>
<tr>
<th>Category</th>
<th>With variants information</th>
<th>Without variants information</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>217 articles</td>
<td>89 articles</td>
<td>306 articles</td>
</tr>
<tr>
<td>Unclear</td>
<td>50 articles</td>
<td>304 articles</td>
<td>354 articles</td>
</tr>
<tr>
<td>Bad</td>
<td>118 articles</td>
<td>1'410 articles</td>
<td>1'528 articles</td>
</tr>
</tbody>
</table>

Tab.6. Results for the medical-annotation dataset.

This table showed that 92% of the Bad articles did not possess any information on variants according to the definition of our patterns. And in 71% of the Good articles, we found at least one of the variants patterns in the titles and abstracts. The variants retrieval appeared to corroborate the articles pre-classification.

The manual review of the 217 Good abstracts in which we found the patterns gave the following results:
- 200 documents were true positives which contained reliable variant information.
- In 13 documents, pattern n°1 retrieved information related to the position in the DNA sequence rather than the amino acid chain.
- In 4 documents, pattern n°4 retrieved a sentence containing a number that was not the position of the amino acid variation.

The results of the review of all Good articles were summarized in the following graph:

Graph.3. Distribution of the Good articles.
The examination of the 89 *Good* articles containing no polymorphisms confirmed that they do not contain any positional information on variants, therefore no false negative. In most of the cases, these articles were about single amino acid polymorphisms but the position of the variation was not expressed in the title and abstract. This information would probably be found in the full-text.

Most of the *Bad* articles were classified as not containing SAPs information. However, there were 118 *Bad* articles that contained at least one of the four patterns. We verified manually these documents; we found:

- 98 articles that were true positives; they effectively contained relevant SAPs.
- 20 articles were false positives generated by the patterns n°1 (9 articles) and n°4 (11 articles).

An example of bad pattern retrieved by the program in an abstract about the zap70 gene (AC P43403) was shown below. The pattern n°4 returned the sentence:

> In contrast, mutations at the tyrosine or leucine residues in the C-terminal YxxL segment abrogate signals for interleukin (IL)-2 production but do not prevent tyrosine phosphorylation of the N-terminal tyrosine of the ITAM, lck association with the ITAM, activation of phospholipase C-gamma1 or calcium mobilization. (PMid 11030411)

This sentence did not describe a polymorphism, but it contained all the required elements: two amino acids tyrosine and leucine, a number 2, and the term mutations. So according to the pattern, we could interpret this sentence as a mutation of a tyrosine into a leucine and vice versa, at position number 2 in the amino acid sequence. This example showed the need for more sophisticated NLP methods for analysing sentences which match with pattern n°4 in order to remove these errors.

To recapitulate, the results for all *Bad* articles of the corpus were shown below:

![Graph 4. Distribution of the *Bad* articles.](image)

The majority, 92% of the *Bad* articles, did not contain information according to our patterns. 6% of documents effectively mentioned variants, but were considered too poor for medical annotation. Only 2% of the corpus was false positive.

For the *Bad* articles that did not contain information on SAPs, we took a sample of 50 articles from the 1'410 in order to verify that they effectively did not contain any information. We
observe no false negatives in this sample. The regular expressions seemed to be able to efficiently reject Bad articles.

Finally the precision of the SAPs retrieval for Good and Bad articles together was calculated. We did not take into account the documents classified as Unclear.

\[
\text{Precision} = \frac{(200 + 98)}{(200 + 98) + (17 + 20)} = 89\%
\]

The precision was 89%, which meant that 89% of the retrieved documents were relevant. The Bad articles were efficiently rejected by the retrieval program, they did not lead to an increase of the number of false positives.

With the Good articles, the recall was 100%, because we had not found any false negatives. Although we could not draw conclusions on so few data, it appeared that in both corpora, there were very few false negatives. The regular expressions appeared to cover well the main different notations of SAPs in publications. Moreover, the articles considered to be Bad were mainly rejected by the program (92%). The regular expressions could thus serve as a strong filter to filter out irrelevant documents in the medical corpora.

### 3.2. SAPs retrieval from PubMed queries

We retrieved a set of articles for each of the 5'222 Swiss-Prot proteins with variants using PubMed query. The query consisted of the name and the synonyms of each of these proteins followed by mutation specific keywords. The articles obtained for the 20 first AC numbers of the human variant protein list (Humpvar.txt, release 51.2, 28 Nov. 2006) were subjected to further analysis. For these 20 proteins, it was found that the number of articles retrieved ranged from 19 to 16'465. In order to facilitate examination, only the first 1’000 articles were analysed if more than 1’000 articles were retrieved from PubMed.

This resulted in a set of 9’116 titles and abstracts for the 20 proteins. Variant terms in these articles were sought using the four regular expressions. The results, protein by protein, are summarized in a table (cf Appendix 4.). On the base of these documents and their extracted SAPs content, we assessed the quality of the query and observed the behaviour of the retrieval program on “real” data.

#### 3.2.1. Pattern matching analysis for the retrieved documents

From the 9’116 publications retrieved above, there were 177 documents (2%) that contained at least one variant and 8’939 documents (98%) in which there was no information based on the patterns. In the 177 abstracts with variant information, we retrieved 380 different variants expressions, distributed as follows:
Contribution of every patterns

46%
15%
3%
36%

Graph. 5. Contribution of every pattern regarding all information extracted.

We observed that the patterns n°1 (46%) and n°4 (36%) were the most important, in agreement with what we observed previously in the Swiss-Prot references and the medical annotation corpora. The total number of articles containing information was much lower (2%) than in the two previously used datasets. This was not surprising as it reflected a real retrieval process on the whole PubMed database.

The 177 matching abstracts were manually reviewed. Problems inherent to the query were found. For example, we observed that while some articles contained the gene or protein name of the protein queried, the protein variants described by the authors belonged to other proteins mentioned in the article.

This was the case with the four proteins named Unknown protein from 2D-PAGE of liver tissue; Spot 94, 13, 34 and 36 (AC P31934, P30091, P30094, and P30096, respectively). For these four proteins, we retrieved from PubMed approximately the same sets of unspecific articles due to their inexplicit names. The SAPs extracted from these documents seem mainly not related to the four “Unknown proteins”.

Other problems we met with gene and protein names were illustrated by the example of the query with Peroxisome biogenesis factor 1. The protein of interest corresponded to the gene PEX1 and was 1’283 residues long (AC O43933). However, there was another Peroxisome biogenesis factor 1, also in Human, coming from the gene PEX2 (PXMP3) which corresponded to a 305 amino acid-long chain (AC P28328). When we looked for documents on the variants of the PEX1 gene product, we retrieved all articles on polymorphisms mentioning Peroxisome biogenesis factor 1 and we obtained all publications on either PEX1 or PEX2 gene products.

We could not circumvent these problems at the query step. Placing more restriction on the query terms would not be efficient in these cases as the errors were intrinsic to the gene or protein names. It would be wiser to use tests to assess the relevance of the SAPs extracted from such documents.

Despite these problems, this trial on 20 proteins showed that the query was, in most cases, sufficiently precise to retrieve relevant variants-related articles.
3.2.2. Positional validation outcome

The positional validation is an important part of the automatic method. It allowed to relate a document to a known variant, or to retrieve references on potential new variants of a protein. To illustrate the procedure, we performed the positional validation process on the SAPs extracted from abstracts related to the 20 first proteins mentioned above. The SAPs retrieved by the regular expressions in every abstract were compared to the list of known variants of the corresponding protein. This was followed by a comparison with the protein sequence in order to discriminate between potential new variants and false positives.

Of the 37’805 articles recuperated by the query on PubMed (without limitation to 1000 documents) for the 20 proteins, only 2’970 articles (~8%) were kept after the validation process. Among these resulting references, approximately 6% of the polymorphisms extracted were already known variants and 94% are potential new variants. The results are detailed in the following table:

<table>
<thead>
<tr>
<th>AC</th>
<th>Name</th>
<th>known variant</th>
<th>potential new variant</th>
<th>Total number of SAPs</th>
<th>number of analysed articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q9UQ90</td>
<td>Paraplegin</td>
<td>0</td>
<td>152</td>
<td>152</td>
<td>8’793</td>
</tr>
<tr>
<td>Q13608</td>
<td>PAF-2</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>164</td>
</tr>
<tr>
<td>O43933</td>
<td>PEX-1</td>
<td>10</td>
<td>17</td>
<td>27</td>
<td>220</td>
</tr>
<tr>
<td>O15381</td>
<td>NVLp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>46</td>
</tr>
<tr>
<td>Q06481</td>
<td>APPL2</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>180</td>
</tr>
<tr>
<td>P05067</td>
<td>APP</td>
<td>140</td>
<td>2’510</td>
<td>2’650</td>
<td>16’465</td>
</tr>
<tr>
<td>Q6YHK3</td>
<td>CD109</td>
<td>0</td>
<td>24</td>
<td>24</td>
<td>254</td>
</tr>
<tr>
<td>P20742</td>
<td>PZP</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>149</td>
</tr>
<tr>
<td>P01023</td>
<td>A2M</td>
<td>14</td>
<td>6</td>
<td>20</td>
<td>609</td>
</tr>
<tr>
<td>P21589</td>
<td>5’-NT</td>
<td>0</td>
<td>13</td>
<td>13</td>
<td>1’004</td>
</tr>
<tr>
<td>Q9H0D6</td>
<td>DHP</td>
<td>0</td>
<td>27</td>
<td>27</td>
<td>861</td>
</tr>
<tr>
<td>P32754</td>
<td>HPD</td>
<td>0</td>
<td>13</td>
<td>13</td>
<td>6’099</td>
</tr>
<tr>
<td>P46952</td>
<td>3-HAO</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>Q15738</td>
<td>NSDHL</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>P26439</td>
<td>HSD II</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>56</td>
</tr>
<tr>
<td>P14060</td>
<td>HSD I</td>
<td>1</td>
<td>10</td>
<td>11</td>
<td>66</td>
</tr>
<tr>
<td>P30096</td>
<td>unknown prot. of 2D PAGE of liver tissue;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>604</td>
</tr>
<tr>
<td></td>
<td>spot 36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P30094</td>
<td>unknown prot. of 2D PAGE of liver tissue;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>549</td>
</tr>
<tr>
<td></td>
<td>spot 34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P30091</td>
<td>unknown prot. of 2D PAGE of liver tissue;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1’328</td>
</tr>
<tr>
<td></td>
<td>spot 13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P31934</td>
<td>unknown prot. of 2D PAGE of liver tissue;</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>313</td>
</tr>
<tr>
<td></td>
<td>spot 94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tab.7. Results of evaluation for 20 proteins.
The four unknown protein of 2D PAGE of liver tissue had been previously identified as a potential source of false positives. Due to their imprecise name, we retrieved several irrelevant articles from PubMed for these proteins. We observed that this bad information was well filtered by the positional validation procedure, as no variants were finally retrieved.

The reading of the context line of the 173 references on known variants showed that all these documents appeared to be reliable, and they mentioned the right SAP. The review of 20 references supposed to contain potential new variants showed that even if their extracted variant corresponds to the amino acid present in the protein sequence, they belong to other proteins mentioned in the article. We found 20 false positives on 20 reviewed references. This showed that there would be a need for manual review of this information or the inclusion of other “filtering” steps to assess the quality of the potential new variants.

3.2.3. Information stored in the database

For the first run of the references retrieval method for all 5'222 proteins (Release 51.4), we counted 2'200'063 possible new references. The whole procedure analysed in average 13 proteins per hour. The stored information was divided as follows:

<table>
<thead>
<tr>
<th>References on known variants</th>
<th>References on potential new variants</th>
<th>Total number of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>23'520</td>
<td>High 470'588</td>
<td>2'200'063</td>
</tr>
<tr>
<td></td>
<td>Medium 1'705'955</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low and very low 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total of new ref. 2'176'543</td>
<td></td>
</tr>
</tbody>
</table>

Tab.8. References repartition for the 5’222 proteins tested.

The references on already known variants were a small part of the retrieved data (1%). The remaining part of the information was composed of references on potential new variants. These documents were mainly classified as having a medium confidence. This category included close to 78% of the retrieved data, the next step would be to apply a more specific sub-classification or filtering to this large set of documents.

The stored information was provided with a context line that allowed quick assessment of the extracted variants. The table 9 shows 3 references found for the SAPs of the alpha-2-macroglobulin (AC P01023). In this example, the context line demonstrated well the relevance of the variants. The two new references on the known variant V1000I might contain additional information on the polymorphism. They were quite recent documents on the A2M variants and their relationships with the Alzheimer disease.

The potential new variant G186T, even if it matched with the A2M sequence, was a polymorphism of the Apo-E protein as clearly shown by the context line. However, the linked article might be interesting as it mentioned a single nucleotide polymorphism G2998A of the A2M. In this case, the reference was not totally uninteresting because even though the extracted variant term was not appropriate, the abstract mentioned the searched protein and one of its SNPs.
The stored data on already known variants needed manual assessment. Although the provided information is “raw”, the data are easily analysed thanks to context line. References on potential new variants must be filtered farther, because the pre-classification is still too general as shown by the elevated number of references in the category “medium”.

4. Discussion

For the first update of the Swiss-Prot references on variant, we parsed several thousands of abstracts to obtain a classified set of variants-related documents. The automatic retrieval system we developed seems to be a useful approach to answer the practical challenges imposed by manual curation. The technique allows one to efficiently deal with numerous text documents and thus to remain informed of the last discoveries. The resulting data supplies potential new data on known or unknown variants, and are an additional information source of phenotypic or physiological consequences of the polymorphism. To update again this information, we could limit the search of articles in time by adding the date of the last run, e.g. Jan. 2007, in the query on PubMed in order to avoid parsing the same documents again.

Although the automatic retrieval allowed one to deal with huge quantity of data, it could not be as accurate as a human curator. Computers could not surmount problems such as the ambiguity in terms used to designate genes, proteins, SNPs and SAPs. The efficiency of the method would be greatly improved by the systematic application of standard nomenclature in publications. Even if efforts were made in this direction, numerous articles did not use the normalized notations of SAPs. Therefore, polymorphisms were expressed in sentences that were difficult to interpret by a program. It was particularly difficult to determine which amino acid was the variant and which number within the sentence represented its position in the protein sequence.

Moreover, still many errors were due to ambiguous gene/protein names which match to our patterns, even if we ignored the matches generated by the queried protein/gene names, in the cases where they had the same kind of notation than SAPs. A solution to these problems would be to use elaborated NLP processes to allow the computer to understand the structure of the sentences and to recognize the gene or protein names in order to avoid confusion with a polymorphism notation. For example, one could use a protein or gene name tagging approach [Franzen and al., 2002]. Tagging gene names would avoid taking SNPs for SAPs. We would...
create a rule saying that if an extracted variant is composed of A, T, G or C and is near to a gene name, it is a SNP rather than a SAP.

This work also pointed out the difficulty to recognize the numbering of the sequence used by the authors. We had chosen to confirm the relevance of the SAPs extracted from text by positional validation based on Swiss-Prot information. Every time the numbering of the position of the variants in Swiss-Prot differed from that of the publication, we calculated potential shifts in the SAPs position that could explain this difference. In several cases, we rescued good SAPs expressions by taking into account, for instance, splicing variants or a loss of a signal/transit sequence. But, this process could lead to false positives as it increased the number of match possibilities and consequently the number of potential mismatches.

We also tolerated a difference in the position of the variant of +/- 1 amino acid with regard to Swiss-Prot numbering, because of the different numbering of the first methionine. Given that the position of the initiator methionine in Swiss-Prot had changed and began at 1 and no more at 0, it would be interesting to evaluate the proportion of errors due to this tolerance, compared to the number of good expressions rescued by this flexibility.

Tolerance of errors and shifts calculation in the position during the validation of potential new variants could explain that the positional validation step filtered only a part of the false positives. Indeed, the references on potential new variants still contained many errors at the end of the whole retrieval process. Moreover, the pre-classification was not sufficiently precise, the main part of the data on new variants were in the medium class. We would improve the classification by defining subclasses based on more specific criteria.

New analysis of these variants would be possible by studying the context of the SAP in text. For instance, the proximity of the variant and the protein name in sentences would inform on the relevance of the variant. This validation technique seemed consistent in view of the fact that it had been used by other research groups that carried out similar works to extract point mutations, for example the tool MuteXt [Horn et al., 2004], Mutation Miner [Baker and al., 2004] and MEMA [Rebholz-Schuhmann and al., 2004]. All these tools aimed to extract point mutations from Medline abstracts. The verification of the extracted information was also based on the position of the variant in the sequence and on advanced NLP procedures measuring the distance in text between a mutation and a gene/protein name. They also brought to light the same difficulties that we met with the nomenclature and the sequence numbering. The most recent tool was MutationGraB [Lee and al., 2007] that aimed to calculate the frequency of terms and the statistical significance of the nearness of the protein or gene name and the mutation in text by using a graph bigram association.

An interesting further improvement of the retrieval process would be to adapt the method to work on full-text articles, in order to retrieve more information. Indeed, in 48% of the Swiss-Prot references corpus, we had not been able to find out polymorphisms in title and abstract, indicating that the information was present somewhere in full-text. However, parsing full-text would require more advanced text-mining technique to determine the structure of the documents in which the variants information would be extracted. This, when combined with advanced NLP procedures for context analysis, would enhance the performance of this automatic retrieval approach for the extraction of new polymorphisms.

5. Conclusion

We have developed an automatic method to retrieve information on variants from the scientific literature. The retrieval procedure was based on four Perl regular expressions representative of the different SAPs notations. The method allowed analyzing on average 13 proteins per hour,
corresponding to approximately 100’000 documents in one hour. The retrieval process had a specificity of more than 84%, and a minimum of recall of 48%. The automatic SAPs retrieval process is efficient to find out new references on known polymorphisms of proteins of interest. The extraction of previous unknown variants would however needs the use of more advanced text-mining and NLP techniques and could be the subject of further development.
Appendix

1. List of terms used in SAPs retrieval program to find out pattern n°4:

Every vocabulary terms can be written in both singular and plural forms and can be capitalize.

- Vocabulary describing polymorphism:
  - *single amino acid polymorphism*, or *polymorphism* alone, or *SAP*,
  - *variant*, *variation*, verbal forms of *to vary*,
  - *substitution*, verbal forms of *to substitute*,
  - *transition*,
  - *transversion*,
  - *point or missense mutation*, verbal forms of *to mutate*,
  - *conversion*, verbal forms of *to convert*,
  - *change*, *exchange* and verbal forms of *to change* or *to exchange*,
  - *replacement* and verbal forms of *to replace*.

- Vocabulary describing position:
  - *amino acid*,
  - *codon*,
  - *position*,
  - *exon*,
  - *localization* or *localisation*, and verbal forms of *to localize* or *to localise*,
  - *situation* and verbal forms of *to situate*,
  - *residue*.

2. List of terms used in SAPs retrieval program for pattern n°4 to know the direction of the variation in the sentence:

Every vocabulary terms can be written in both singular and plural forms and can be capitalize.

- Vocabulary describing the positional terms, if a number is found previous or following these terms, we regard it as the position of the SAP in the protein sequence:
  - *amino acid*,
  - *position*,
  - *localization* or *localisation*, verbal forms *to localize* or *to localise*,
  - *residue*,
  - verbal forms of *to situate*

For instance: “Point mutation at residue 345 changing isoleucine to valine”, the position of the mutation.

- Preposition indicating the sense of the amino acid change:
  - *to*, *into*, or *by*.

In the sentence: “Point mutation at residue 345 changing isoleucine to valine”, the firstly mentioned amino acid is the one we found normally in the sequence and the amino acid mentioned after the preposition *to* is the one that is replacing it.
3. List of expressions that lead to false positives with pattern n°1. The context terms are the terms in the context sentences that allow recognizing these misleading expressions:

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Context terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>G2M and G1S</td>
<td>- cell cycle,</td>
</tr>
<tr>
<td></td>
<td>- cell division,</td>
</tr>
<tr>
<td></td>
<td>- cycle arrest,</td>
</tr>
<tr>
<td></td>
<td>- cycle phase,</td>
</tr>
<tr>
<td></td>
<td>- mitosis,</td>
</tr>
<tr>
<td></td>
<td>- mitotic transition.</td>
</tr>
<tr>
<td>E2F</td>
<td>- (epidermal) growth factor,</td>
</tr>
<tr>
<td></td>
<td>- transcription or transcription factor</td>
</tr>
<tr>
<td>T47D, C13T and C3H</td>
<td>- cell,</td>
</tr>
<tr>
<td></td>
<td>- cell line,</td>
</tr>
<tr>
<td></td>
<td>- mouse or mice.</td>
</tr>
<tr>
<td>H2A</td>
<td>- histone (2A)</td>
</tr>
<tr>
<td>M24R</td>
<td>- Brandel Laboratories (Gaithersburg, MD),</td>
</tr>
<tr>
<td></td>
<td>- filter,</td>
</tr>
<tr>
<td></td>
<td>- cell harvester.</td>
</tr>
</tbody>
</table>

4. Content of the tables in which we stored potential new references:

The polymorphisms are stored in the ‘Polymorphism table’. The corresponding references are stored separately in the ‘Reference table’ to avoid redundancy as a document may cite many different SAPs and a SAP may possess many different references.

The PMIDs are the keys to link the polymorphisms table to the references table. The design of the tables is shown below:

**Polymorphisms table**

<table>
<thead>
<tr>
<th>Isold</th>
<th>position</th>
<th>from aa</th>
<th>to aa</th>
<th>PPLd</th>
<th>PMID</th>
<th>score</th>
<th>context</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC-1</td>
<td></td>
<td>1 letter amino acid code</td>
<td>1 letter amino acid code</td>
<td>foreign key</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**References table**

<table>
<thead>
<tr>
<th>PMID</th>
<th>Authors</th>
<th>Title</th>
<th>Reference location</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary key</td>
<td></td>
<td></td>
<td>Journal; volume; pages; year;</td>
</tr>
</tbody>
</table>
The polymorphisms table contains:
- IsoId: unique identifier of the protein, composed by the Swiss-Prot AC number, followed by a hyphen, followed by 1.
- The position of the polymorphism in the amino acid sequence.
- The amino acid present in the reference protein.
- The polymorphic amino acid.
- The FTId, unique identifier of the already reported variants; this field is null for potential new variants.
- The context, it means the sentence containing the variants expression (~20 words).
- A score of confidence for new variants (high = 3, medium = 2, low = 1, very low = 0), maximal for references on known variants (known variant = 4).
- PMId of the article mentioning the variant.

The references table contains:
- PMId of the article,
- Authors,
- Title,
- Reference location; it means journal, volume, pages, and year of publication.
5. Results of the SAPs retrieval program in articles from PubMed for 20 proteins:

Maximal size of the corpus for each AC is 1’000 articles.

<table>
<thead>
<tr>
<th>Protein and gene names used to query PubMed</th>
<th>AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown protein from 2D-PAGE of liver tissue; Spot 94;</td>
<td>P31934</td>
</tr>
<tr>
<td>Unknown protein from 2D-PAGE of plasma; Spot 13;</td>
<td>P30091</td>
</tr>
<tr>
<td>Unknown protein from 2D-PAGE of plasma; Spot 34;</td>
<td>P30094</td>
</tr>
<tr>
<td>Unknown protein from 2D-PAGE of plasma; Spot 36;</td>
<td>P30096</td>
</tr>
<tr>
<td>3 beta-hydroxysteroid dehydrogenase/Delta 5--&gt;4-isomerase type I; 3-beta-HSD I; Trophoblast antigen FDO161G; HSD3B1; 3BH; HSDB3A;</td>
<td>P14060</td>
</tr>
<tr>
<td>3 beta-hydroxysteroid dehydrogenase/Delta 5---&gt;4-isomerase type II; 3-beta-HSD II; HSD3B2; HSDB3B;</td>
<td>P26439</td>
</tr>
<tr>
<td>Sterol-4-alpha-carboxylate 3-dehydrogenase, decarboxylating; H105e3 protein; NSDHL; H105E3;</td>
<td>Q15738</td>
</tr>
<tr>
<td>3-hydroxyanthranilic acid 3-dehydrogenase; 3-HAO; 3-hydroxyanthranilic acid dioxygenase; 3-hydroxyanthranilic acid dioxygenase; HAAO;</td>
<td>P46952</td>
</tr>
<tr>
<td>4-hydroxyphenylpyruvate dioxygenase; 4HPPDase; 4-hydroxyphenylpyruvic acid oxidase; HPD; PPD;</td>
<td>P32754</td>
</tr>
<tr>
<td>5'-3' exoribonuclease 2; DHM1-like protein; DHP protein; XRN2;</td>
<td>Q9H0D6</td>
</tr>
<tr>
<td>5'-nucleotidase precursor; Ecto-5'-nucleotidase; 5'-NT; CD73 antigen; NT5E; NT5;</td>
<td>P21589</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>566</td>
</tr>
<tr>
<td>3</td>
<td>146</td>
</tr>
<tr>
<td>5</td>
<td>249</td>
</tr>
<tr>
<td>13</td>
<td>987</td>
</tr>
<tr>
<td>7</td>
<td>173</td>
</tr>
<tr>
<td>0</td>
<td>46</td>
</tr>
<tr>
<td>15</td>
<td>205</td>
</tr>
<tr>
<td>7</td>
<td>157</td>
</tr>
<tr>
<td>5</td>
<td>995</td>
</tr>
</tbody>
</table>

Total = 177 articles
Total = 8939 articles
Bibliography

These:


Tools:


Unpublished document:


Articles:


Gajendran VK, Lin JR, Fyhrie DP. An application of bioinformatics and text mining to the discovery of novel genes related to bone biology. Bone. 2007 Jan 20;


Pearson PL. Historical development of analysing large-scale changes in the human genome. Cytogenet Genome Res. 2006;115(3-4):198-204.


