Summarizing Biological Literature with BioSumm

Elena Baralis, Alessandro Fiori
Politecnico di Torino

The availability of increasingly wider text repositories requires effective techniques to manage the huge mass of unstructured information there contained (e.g., navigate, analyze and represent it in the most suitable way). Particularly, in the biological and biomedical domain a huge amount of information is daily generated and contributed by a vast research community spread all over the world. Repositories like PubMed Central, the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature, nowadays contain billions of documents.

System Architecture

- ✓ Fully modular
- ✓ Allows integrating plugins addressed to a specific task (e.g., clustering, web search).
- ✓ Change Domain dictionary to tailor the grading function according to the application domain

Online search & local repository. Keyword search of scientific papers on:
- *Google Scholar,*
- *PubMed Central (PMC)*
- *PubMed*

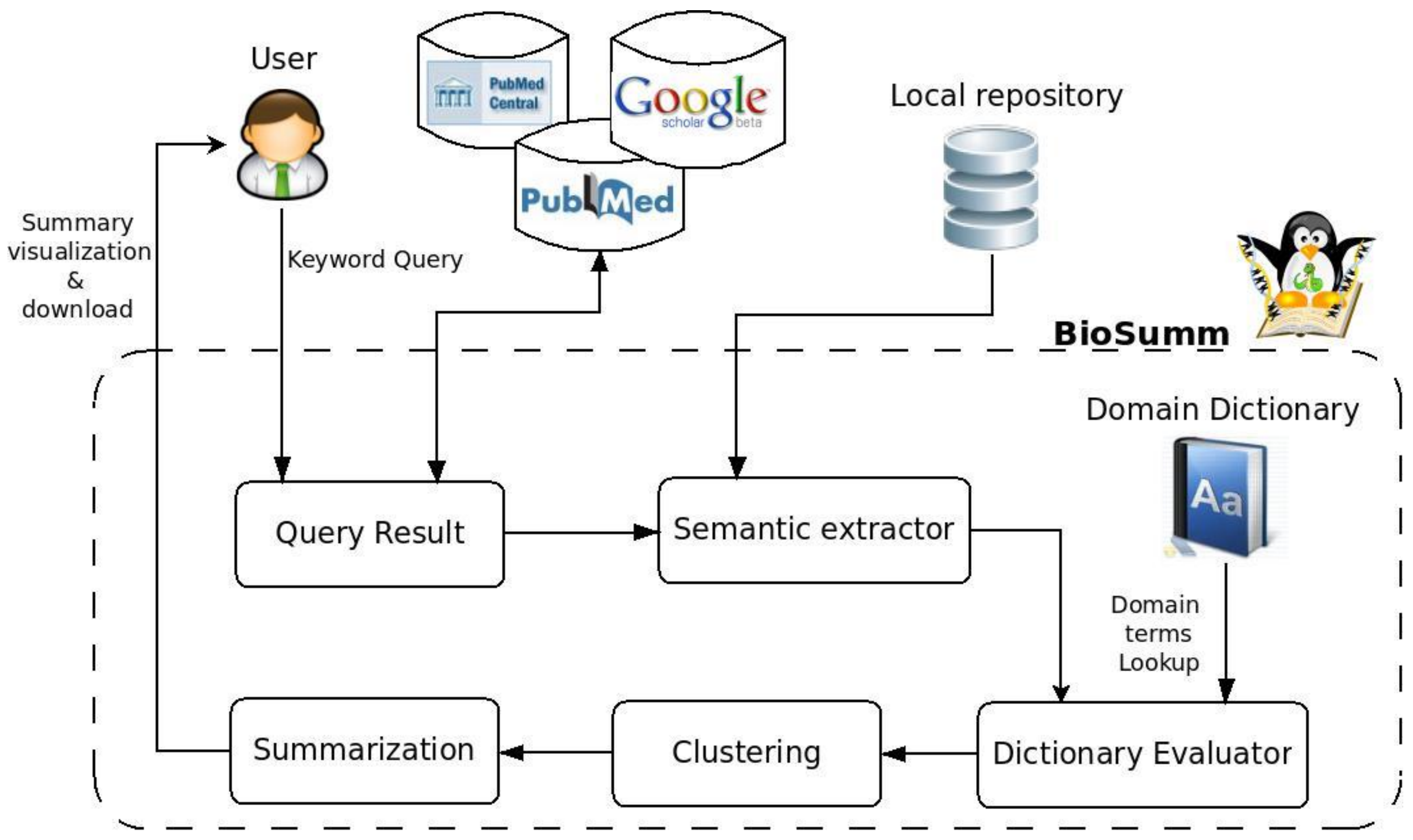
Or load locally stored documents in pdf and xml formats.

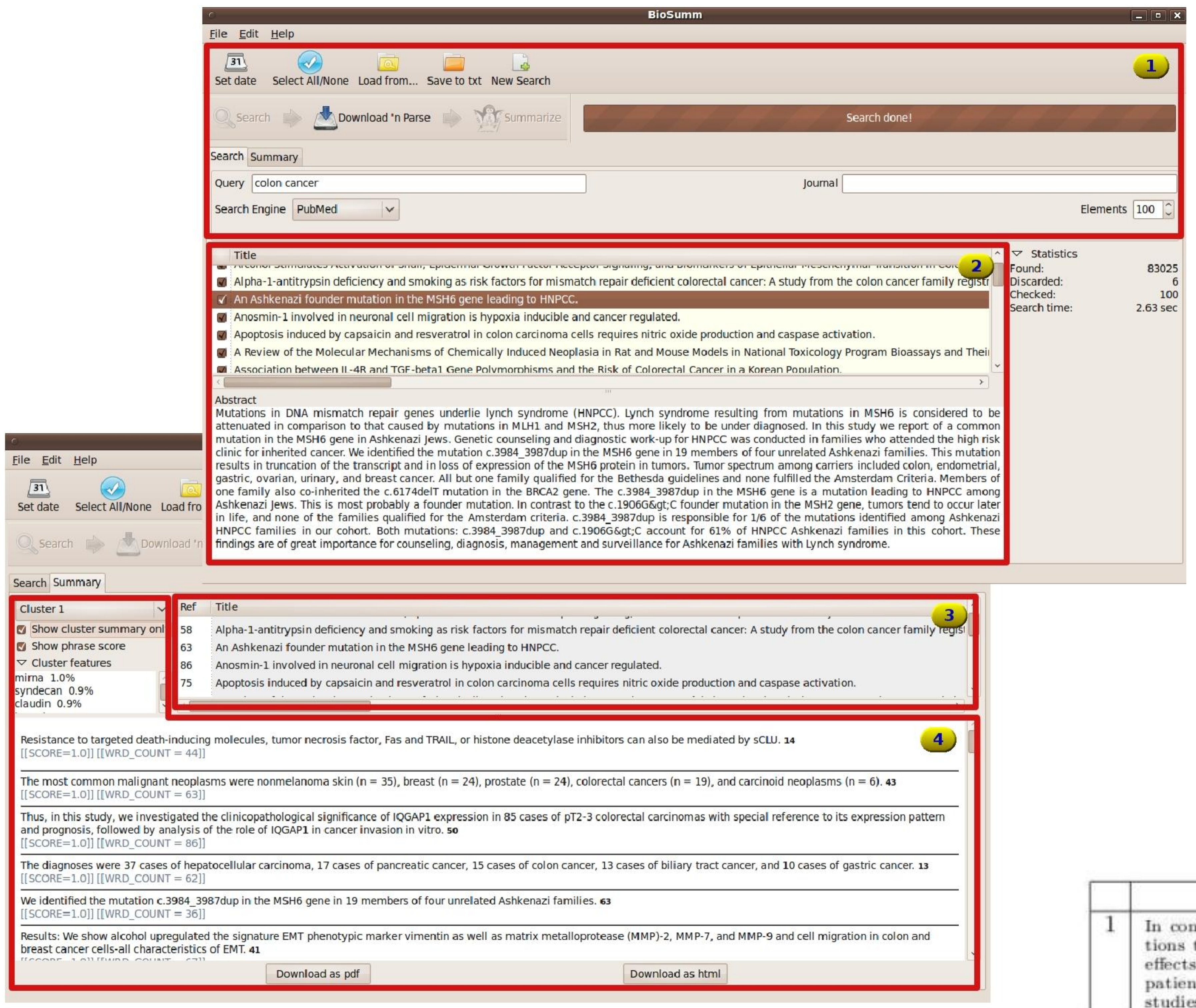
Document structure extractor. Parsing and extraction of available components (e.g., title, authors, journal, abstract, body, keywords).

Clustering. Reduce the heterogeneity of the retrieved documents.
- Document collection represented as a matrix tf-idf
- ***Bisecting K-means*** clustering method is performed

Dictionary evaluator. Compute the semantic weights of grading function according to the terms in the domain dictionary. The Domain dictionary contains genes and proteins names from ***BioGrid*** database.

Summarization. Based on a traditional statistic summarizer, it biases sentence selection using the information contained in the ***Domain Dictionary.***





Search done!

Query: colon cancer
Journal:
Search Engine: PubMed
Elements: 100

Title
- ☑ Alpha-1-antitrypsin deficiency and smoking as risk factors for mismatch repair deficient colorectal cancer: A study from the colon cancer family registr
- ☑ An Ashkenazi founder mutation in the MSH6 gene leading to HNPCC.
- ☑ Anosmin-1 involved in neuronal cell migration is hypoxia inducible and cancer regulated.
- ☑ Apoptosis induced by capsaicin and resveratrol in colon carcinoma cells requires nitric oxide production and caspase activation.
- ☑ A Review of the Molecular Mechanisms of Chemically Induced Neoplasia in Rat and Mouse Models in National Toxicology Program Bioassays and Thei
- ☑ Association between IL-4R and TGF-beta1 Gene Polymorphisms and the Risk of Colorectal Cancer in a Korean Population.

Statistics
Found: 83025
Discarded: 6
Checked: 100
Search time: 2.63 sec

Abstract
Mutations in DNA mismatch repair genes underlie lynch syndrome (HNPCC). Lynch syndrome resulting from mutations in MSH6 is considered to be attenuated in comparison to that caused by mutations in MLH1 and MSH2, thus more likely to be under diagnosed. In this study we report of a common mutation in the MSH6 gene in Ashkenazi Jews. Genetic counseling and diagnostic work-up for HNPCC was conducted in families who attended the high risk clinic for inherited cancer. We identified the mutation c.3984_3987dup in the MSH6 gene in 19 members of four unrelated Ashkenazi families. This mutation results in truncation of the transcript and in loss of expression of the MSH6 protein in tumors. Tumor spectrum among carriers included colon, endometrial, gastric, ovarian, urinary, and breast cancer. All but one family qualified for the Bethesda guidelines and none fulfilled the Amsterdam Criteria. Members of one family also co-inherited the c.6174delT mutation in the BRCA2 gene. The c.3984_3987dup in the MSH6 gene is a mutation leading to HNPCC among Ashkenazi Jews. This is most probably a founder mutation. In contrast to the c.1906G&gt;C founder mutation in the MSH2 gene, tumors tend to occur later in life, and none of the families qualified for the Amsterdam criteria. c.3984_3987dup is responsible for 1/6 of the mutations identified among Ashkenazi HNPCC families in our cohort. Both mutations: c.3984_3987dup and c.1906G&gt;C account for 61% of HNPCC Ashkenazi families in this cohort. These findings are of great importance for counseling, diagnosis, management and surveillance for Ashkenazi families with Lynch syndrome.

Cluster 1
- ☑ Show cluster summary onl
- ☑ Show phrase score
- Cluster features

mirna	1.0%
syndecan	0.9%
claudin	0.9%

Ref	Title
58	Alpha-1-antitrypsin deficiency and smoking as risk factors for mismatch repair deficient colorectal cancer: A study from the colon cancer family regist
63	An Ashkenazi founder mutation in the MSH6 gene leading to HNPCC.
86	Anosmin-1 involved in neuronal cell migration is hypoxia inducible and cancer regulated.
75	Apoptosis induced by capsaicin and resveratrol in colon carcinoma cells requires nitric oxide production and caspase activation.

Resistance to targeted death-inducing molecules, tumor necrosis factor, Fas and TRAIL, or histone deacetylase inhibitors can also be mediated by sCLU. **14**
[[SCORE=1.0]] [[WRD_COUNT = 44]]

The most common malignant neoplasms were nonmelanoma skin (n = 35), breast (n = 24), prostate (n = 24), colorectal cancers (n = 19), and carcinoid neoplasms (n = 6). **43**
[[SCORE=1.0]] [[WRD_COUNT = 63]]

Thus, in this study, we investigated the clinicopathological significance of IQGAP1 expression in 85 cases of pT2-3 colorectal carcinomas with special reference to its expression pattern and prognosis, followed by analysis of the role of IQGAP1 in cancer invasion in vitro. **50**
[[SCORE=1.0]] [[WRD_COUNT = 86]]

The diagnoses were 37 cases of hepatocellular carcinoma, 17 cases of pancreatic cancer, 15 cases of colon cancer, 13 cases of biliary tract cancer, and 10 cases of gastric cancer. **13**
[[SCORE=1.0]] [[WRD_COUNT = 62]]

We identified the mutation c.3984_3987dup in the MSH6 gene in 19 members of four unrelated Ashkenazi families. **63**
[[SCORE=1.0]] [[WRD_COUNT = 36]]

Results: We show alcohol upregulated the signature EMT phenotypic marker vimentin as well as matrix metalloprotease (MMP)-2, MMP-7, and MMP-9 and cell migration in colon and breast cancer cells-all characteristics of EMT. **41**

Graphical User Interface

1. **Document search.** The user can set the parameters (e.g., keywords, journal, date of publication) to retrieve the documents from supported digital libraries.

2. **Document browsing.** Management of retrieved documents to select the most relevant for summarization task. Documents may be individually excluded from the document collection on which both clustering and summarization are performed, thus further refining the analysis.

3. **Documents of cluster.** List of the documents belonging to a cluster identified by the clustering block. BibTex description is also available.

4. **Cluster summary.** Each sentence in the summary is scored by relevance with respect to the considered application domain. Furthermore, the cluster is described by means of the relevant keywords appearing in its documents.

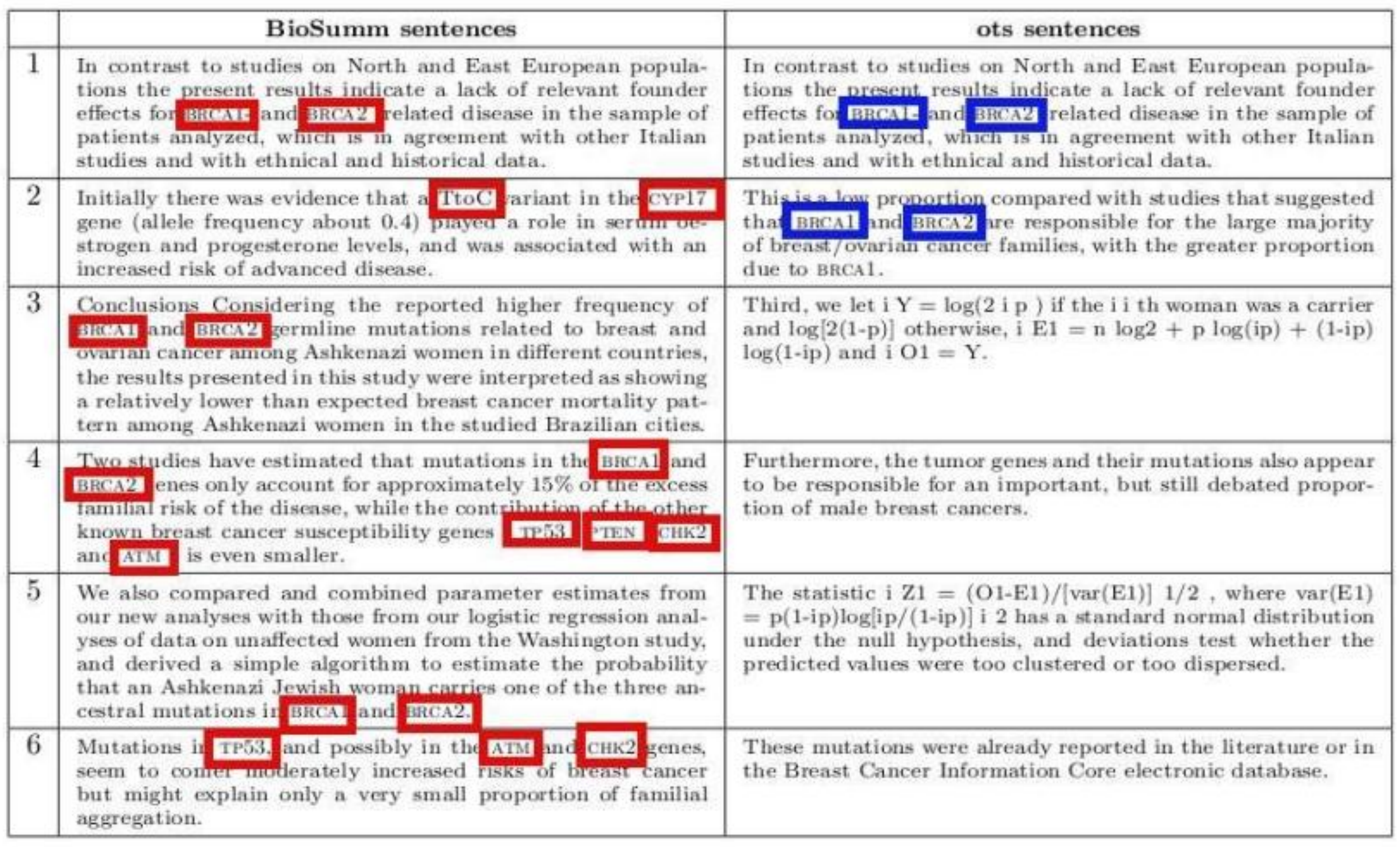
Experimental results

BioSumm is neither a traditional summarizer nor a extractor of dictionary terms. It is designed to be a summarizer oriented to the biological domain. Thus, its summaries have both the expressive power of the traditional summaries and the domain specificity of documents produced by a dictionary entry extractor.

The difference with a traditional summarizer may be appreciated in the example. Six most graded sentences in BioSumm and in a traditional summary are reported. The results were produced by the experiments carried on the scientific journals freely available in *PubMed Central*. Specifically, they contain sentences belonging to a cluster of documents belonging to the Breast Cancer journal. The keywords of the cluster (the words describing its major topics) are *proband, Ashkenazi, Jewish* .

The comparison shows that BioSumm, although oriented on *biology*, is still able to cover all the major topics covered by a traditional summarizer. Moreover, its sentences are less generic and contains a lot of genes and proteins which are described in details and not only listed.

The results suggest that researchers that discover gene correlations by means of analysis tools (e.g., data mining tools) may exploit this framework to effectively support the biological validation of their results. Experimental results obtained by means of *ROUGE* are also reported.



	BioSumm sentences	ots sentences
1	In contrast to studies on North and East European populations the present results indicate a lack of relevant founder effects for BRCA1 and BRCA2 related disease in the sample of patients analyzed, which is in agreement with other Italian studies and with ethnical and historical data.	In contrast to studies on North and East European populations the present results indicate a lack of relevant founder effects for BRCA1 and BRCA2 related disease in the sample of patients analyzed, which is in agreement with other Italian studies and with ethnical and historical data.
2	Initially there was evidence that a TtoC variant in the CYP17 gene (allele frequency about 0.4) played a role in serum oestrogen and progesterone levels, and was associated with an increased risk of advanced disease.	This is a low proportion compared with studies that suggested that BRCA1 and BRCA2 are responsible for the large majority of breast/ovarian cancer families, with the greater proportion due to BRCA1.
3	Conclusions Considering the reported higher frequency of BRCA1 and BRCA2 germline mutations related to breast and ovarian cancer among Ashkenazi women in different countries, the results presented in this study were interpreted as showing a relatively lower than expected breast cancer mortality pattern among Ashkenazi women in the studied Brazilian cities.	Third, we let i Y = log(2 i p) if the i i th woman was a carrier and log[2(1-p)] otherwise, i E1 = n log2 + p log(ip) + (1-ip) log(1-ip) and i O1 = Y.
4	Two studies have estimated that mutations in the BRCA1 and BRCA2 genes only account for approximately 15% of the excess familial risk of the disease, while the contribution of the other known breast cancer susceptibility genes TP53, PTEN, CHK2 and ATM is even smaller.	Furthermore, the tumor genes and their mutations also appear to be responsible for an important, but still debated proportion of male breast cancers.
5	We also compared and combined parameter estimates from our new analyses with those from our logistic regression analyses of data on unaffected women from the Washington study, and derived a simple algorithm to estimate the probability that an Ashkenazi Jewish woman carries one of the three ancestral mutations in BRCA1 and BRCA2.	The statistic i Z1 = (O1-E1)/[var(E1)] 1/2 , where var(E1) = p(1-ip)log[ip/(1-ip)] i 2 has a standard normal distribution under the null hypothesis, and deviations test whether the predicted values were too clustered or too dispersed.
6	Mutations in TP53, and possibly in the ATM and CHK2 genes, seem to confer moderately increased risks of breast cancer but might explain only a very small proportion of familial aggregation.	These mutations were already reported in the literature or in the Breast Cancer Information Core electronic database.

Experimental comparison between BioSumm and general purpose summarizer (OTS)

Dataset	BioSumm			OTS		
	Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
Breast Cancer	**0.08246**	**0.22553**	**0.11456**	0.08026	0.21860	0.11141
Arthritis Res	**0.09089**	**0.25362**	**0.12596**	0.08844	0.24406	0.12197

ROUGE-2 evaluation on two different document collections

Dataset	BioSumm			OTS		
	Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
Breast Cancer	**0.10038**	**0.28175**	**0.14053**	0.09872	0.27599	0.1811
Arthritis Res	**0.11095**	**0.31777**	**0.15498**	0.10905	0.30888	0.15169

ROUGE-SU4 evaluation on two different document collections

Contact info
Alessandro Fiori (Post-Doc Researcher) - Phone: +39 011 090 7194 - Fax: +39 011 090 7099 - Email: alessandro.fiori@ polito.it -Web: http://dbdmg.polito.it/