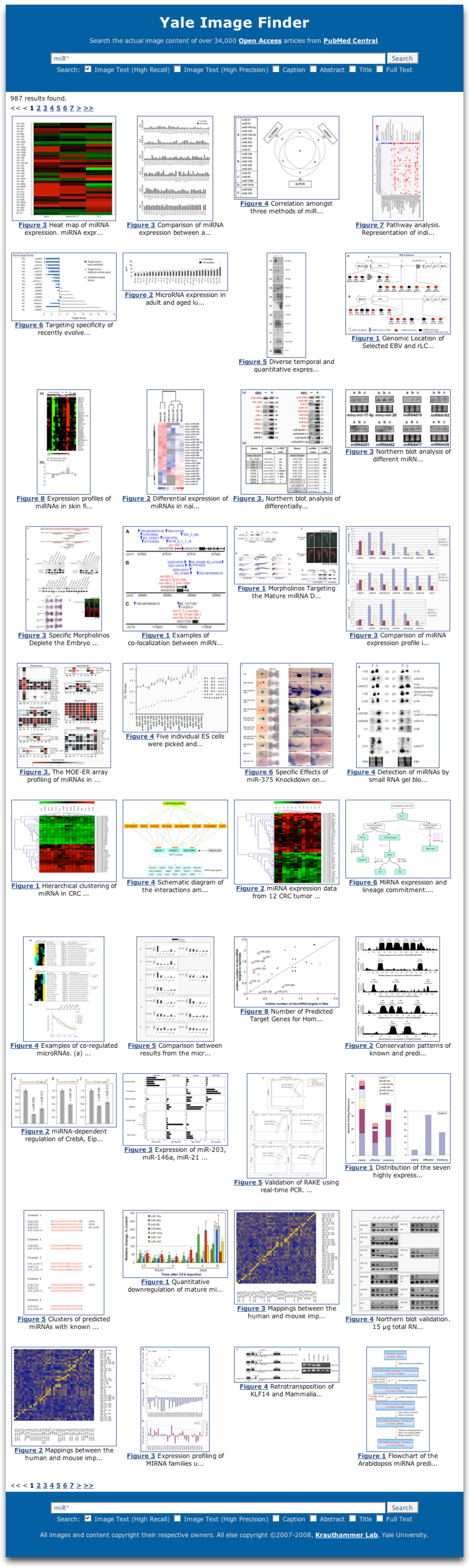
Using textual context for improving OCR performance in biomedical literature retrieval

Songhua Xu[1], James McCusker[2], Martin Schultz[1], and Michael Krauthammer[2]
[1]Department of Computer Science, [2]Department of Pathology and Yale Center for Medical Informatics, Yale University, New Haven, CT

http://krauthammerlab.med.yale.edu/imagefinder

Abstract

Today's information retrieval (IR) techniques are mostly text-based, which fail in situations when textual information is not easily accessible, such as in biomedical images and figures. We propose to augment IR with optical character recognition (OCR) capabilities, and describe a context-based method for boosting OCR performance.



Yale Image Finder

Search the actual image content of over 34,000 Open Access articles from PubMed Central.

miR* | Search

Search: ☑ Image Text (High Recall) □ Image Text (High Precision) □ Caption □ Abstract □ Title □ Full Text

987 results found.

<< < 1 2 3 4 5 6 7 > >>

Figure 3 Heat map of miRNA expression. miRNA expr...

Figure 3 Comparison of miRNA expression between a...

Figure 4 Correlation amongst three methods of miR...

Figure 7 Pathway analysis. Representation of indi...

Figure 6 Targeting specificity of recently evolve...

Figure 2 MicroRNA expression in adult and aged lu...

Figure 5 Diverse temporal and quantitative expres...

Figure 1 Genomic Location of Selected EBV and rLC...

Figure 8 Expression profiles of miRNAs in skin fi...

Figure 2 Differential expression of miRNAs in naï...

Figure 3. Northern blot analysis of differentially...

Figure 3 Northern blot analysis of different miRN...

Figure 3 Specific Morpholinos Deplete the Embryo ...

Figure 1 Examples of co-localization between miRN...

Figure 1 Morpholinos Targeting the Mature miRNA D...

Figure 3 Comparison of miRNA expression profile i...

Figure 3. The MOE-ER array profiling of miRNAs in ...

Figure 4 Five individual ES cells were picked and...

Figure 6 Specific Effects of miR-375 Knockdown on...

Figure 4 Detection of miRNAs by small RNA gel blo...

Figure 1 Hierarchical clustering of miRNA in CRC ...

Figure 4 Schematic diagram of the interactions am...

Figure 2 miRNA expression data from 12 CRC tumor ...

Figure 6 MiRNA expression and lineage commitment....

Figure 4 Examples of co-regulated microRNAs. (a) ...

Figure 5 Comparison between results from the micr...

Figure 8 Number of Predicted Target Genes for Hom...

Figure 2 Conservation patterns of known and predi...

Figure 2 miRNA-dependent regulation of CrebA, Eip...

Figure 3 Expression of miR-203, miR-146a, miR-21 ...

Figure 5 Validation of RAKE using real-time PCR. ...

Figure 1 Distribution of the seven highly express...

Figure 5 Clusters of predicted miRNAs with known ...

Figure 1 Quantitative downregulation of mature mi...

Figure 3 Mappings between the human and mouse imp...

Figure 4 Northern blot validation. 15 µg total RN...

Figure 2 Mappings between the human and mouse imp...

Figure 3 Expression profiling of MIRNA families u...

Figure 4 Retrotransposition of KLF14 and Mammalia...

Figure 1 Flowchart of the Arabidopsis miRNA predi...

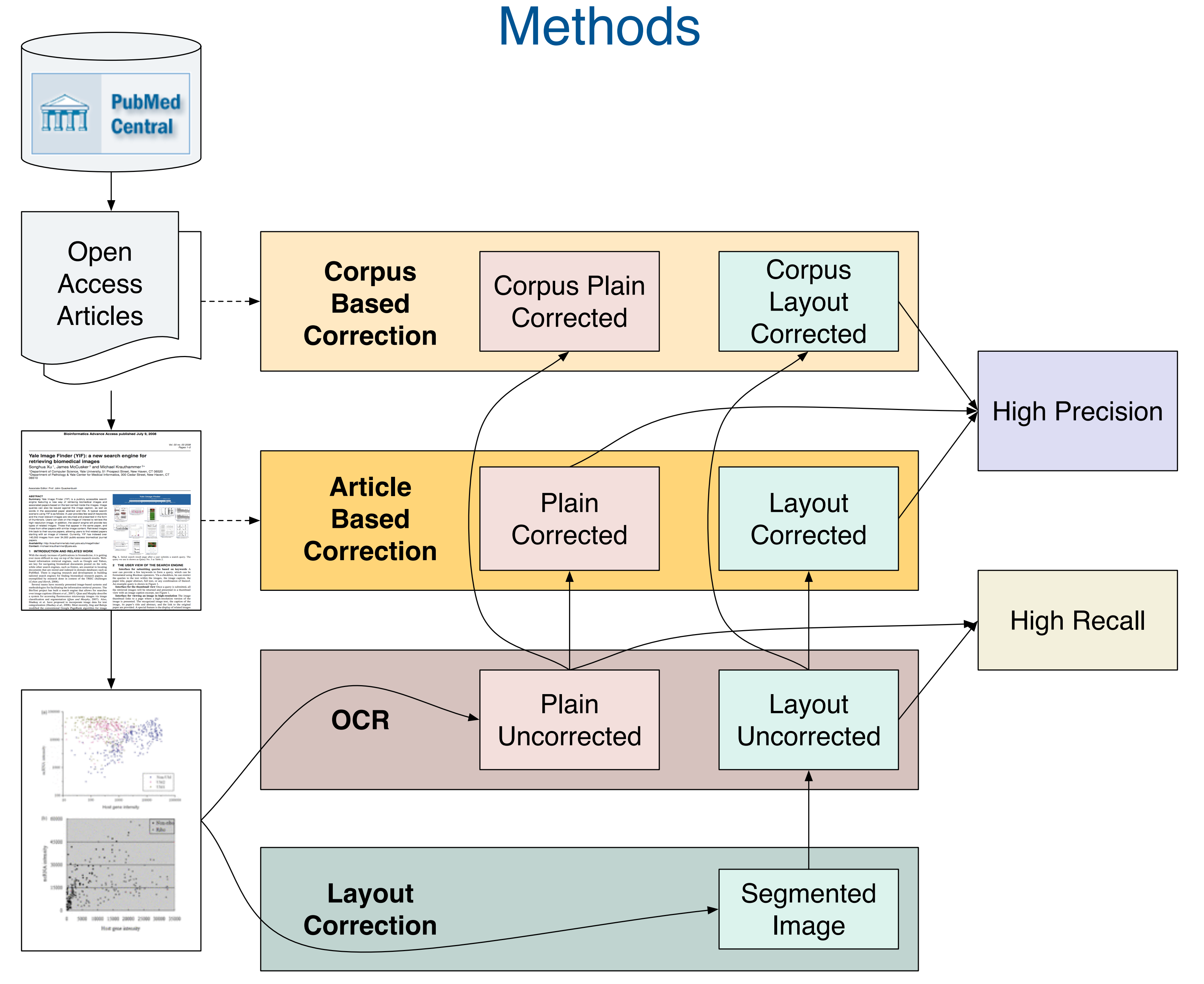
<< < 1 2 3 4 5 6 7 > >>

miR* | Search

Search: ☑ Image Text (High Recall) □ Image Text (High Precision) □ Caption □ Abstract □ Title □ Full Text

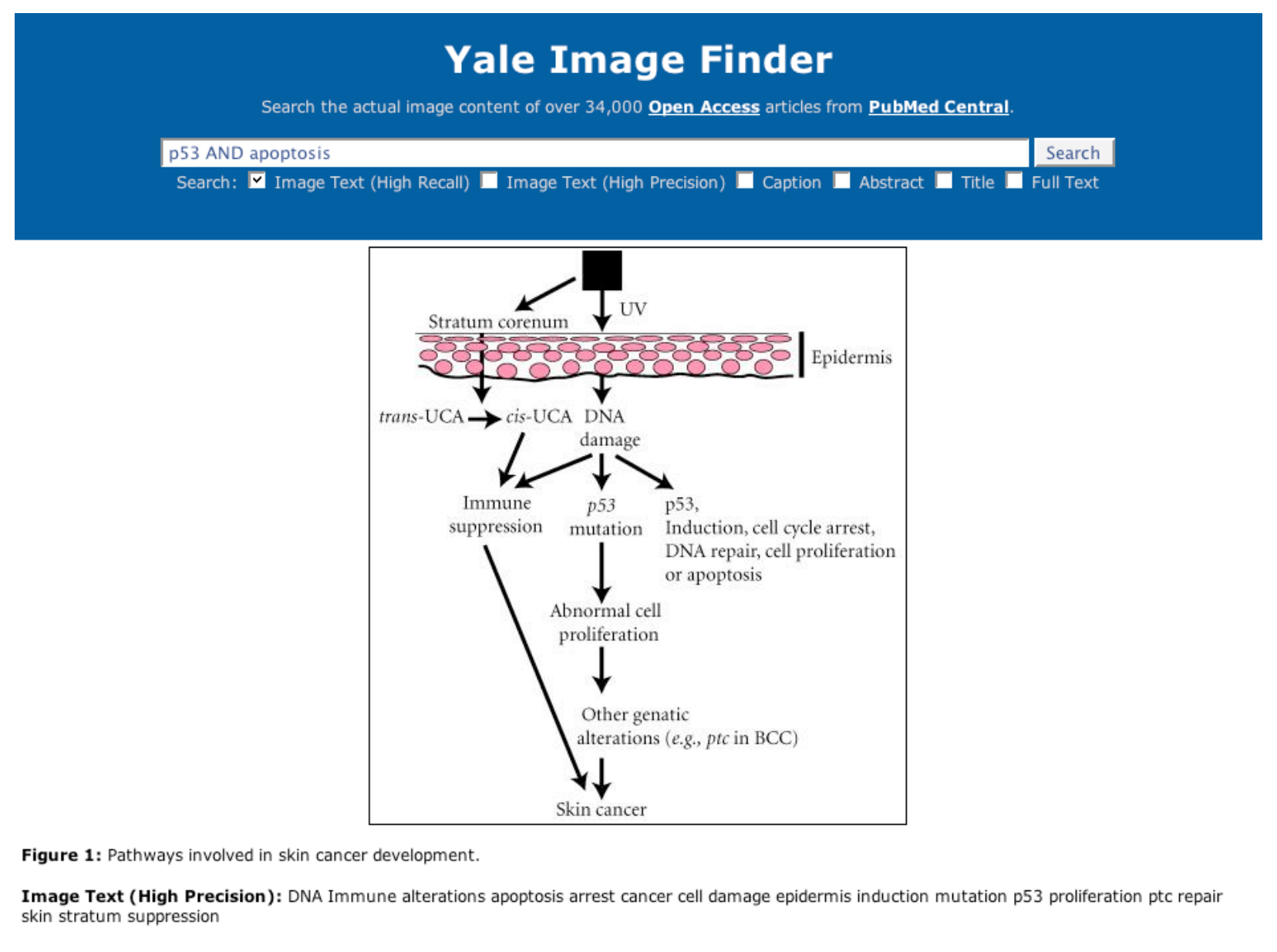
Yale Image Finder search results for "miR*" in high-recall mode.

Methods



Advantages of OCR-based image text extraction

	Query	Search Target Domain	Graph	Gel	Microscopy	Diagram	List	Misc.	Total		Relevant		Precision
1	diet AND insulin	Caption	17	0	0	0	0	2	19		19		100%
		Caption & Image Text (HR)	17	1	0	6	0	3	27		25		93%
		Δ	0	1	0	6	0	1	8	42%	6	32%	75%
2	apoptosis AND p53	Caption	11	1	5	11	0	14	42		42		100%
		Caption & Image Text (HR)	12	1	5	18	4	15	55		54		98%
		Δ	1	0	0	7	4	1	13	31%	12	29%	92%
3	miR* AND brain AND heart	Caption	1	1	0	0	1	2	5		4		80%
		Caption & Image Text (HR)	1	3	0	0	6	3	13		11		85%
		Δ	0	2	0	0	5	1	8	160%	7	175%	88%



Yale Image Finder

Search the actual image content of over 34,000 Open Access articles from PubMed Central.

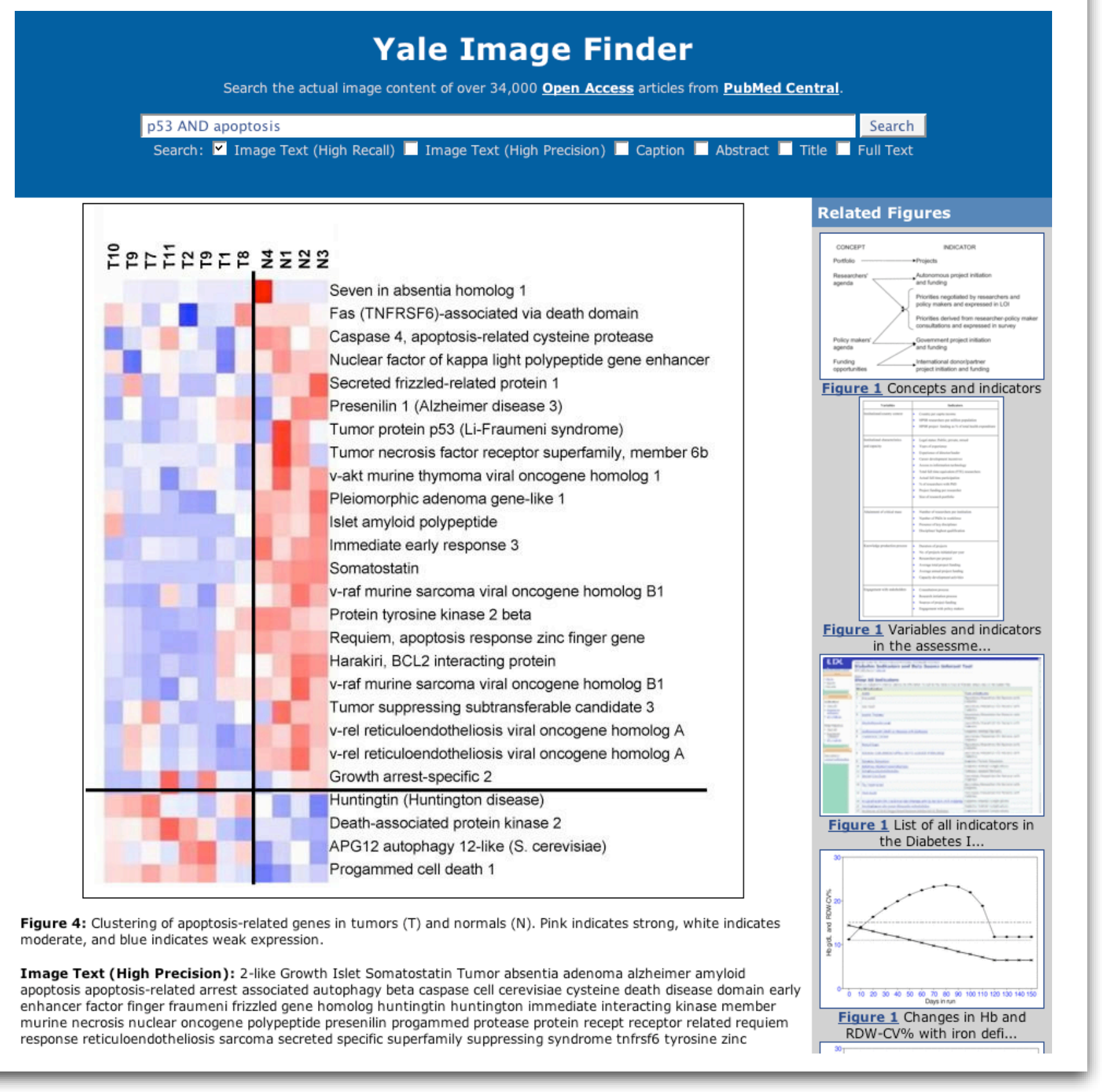
p53 AND apoptosis | Search

Search: ☑ Image Text (High Recall) □ Image Text (High Precision) □ Caption □ Abstract □ Title □ Full Text

UV
Stratum corneum
Epidermis
trans-UCA
cis-UCA
DNA damage
Immune suppression
p53 mutation
p53, Induction, cell cycle arrest, DNA repair, cell proliferation or apoptosis
Abnormal cell proliferation
Other genatic alterations (e.g., ptc in BCC)
Skin cancer

Figure 1: Pathways involved in skin cancer development.

Image Text (High Precision): DNA Immune alterations apoptosis arrest cancer cell damage epidermis induction mutation p53 proliferation ptc repair skin stratum suppression



Yale Image Finder

Search the actual image content of over 34,000 Open Access articles from PubMed Central.

p53 AND apoptosis | Search

Search: ☑ Image Text (High Recall) □ Image Text (High Precision) □ Caption □ Abstract □ Title □ Full Text

T10
T9
T7
T11
T2
T9
T1
T8
N4
N1
N2
N3
Seven in absentia homolog 1
Fas (TNFRSF6)-associated via death domain
Caspase 4, apoptosis-related cysteine protease
Nuclear factor of kappa light polypeptide gene enhancer
Secreted frizzled-related protein 1
Presenilin 1 (Alzheimer disease 3)
Tumor protein p53 (Li-Fraumeni syndrome)
Tumor necrosis factor receptor superfamily, member 6b
v-akt murine thymoma viral oncogene homolog 1
Pleiomorphic adenoma gene-like 1
Islet amyloid polypeptide
Immediate early response 3
Somatostatin
v-raf murine sarcoma viral oncogene homolog B1
Protein tyrosine kinase 2 beta
Requiem, apoptosis response zinc finger gene
Harakiri, BCL2 interacting protein
v-raf murine sarcoma viral oncogene homolog B1
Tumor suppressing subtransferable candidate 3
v-rel reticuloendotheliosis viral oncogene homolog A
v-rel reticuloendotheliosis viral oncogene homolog A
Growth arrest-specific 2
Huntingtin (Huntington disease)
Death-associated protein kinase 2
APG12 autophagy 12-like (S. cerevisiae)
Progammed cell death 1

Figure 4: Clustering of apoptosis-related genes in tumors (T) and normals (N). Pink indicates strong, white indicates moderate, and blue indicates weak expression.

Image Text (High Precision): 2-like Growth Islet Somatostatin Tumor absentia adenoma alzheimer amyloid apoptosis apoptosis-related arrest associated autophagy beta caspase cell cerevisiae cysteine death disease domain early enhancer factor finger fraumeni frizzled gene homolog huntingtin huntington immediate interacting kinase member murine necrosis nuclear oncogene polypeptide presenilin progammed protease protein recept receptor related requiem response reticuloendotheliosis sarcoma secreted specific superfamily suppressing syndrome tnfrsf6 tyrosine zinc

Related Figures

Figure 1 Concepts and indicators

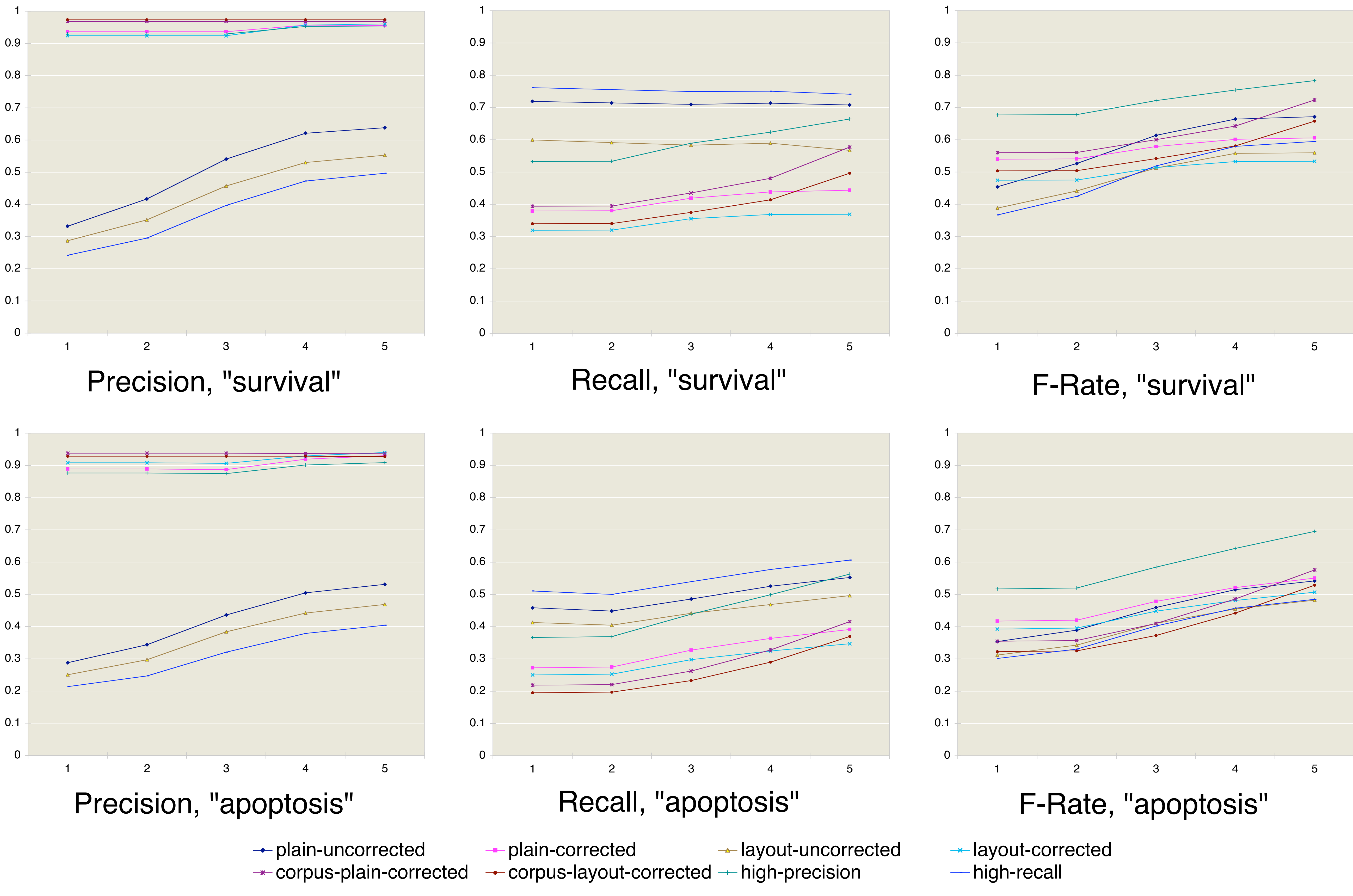
Figure 1 Variables and indicators in the assessme...

Figure 1 List of all indicators in the Diabetes I...

Figure 1 Changes in Hb and RDW-CV% with iron defi...

Figures with short or insufficient captions are often missed by caption-based search. A combination of captions and image text offers the greatest range of searchable text. This is especially true in diagrams, lists, and heatmaps.

Evaluation



Precision, "survival"

Recall, "survival"

F-Rate, "survival"

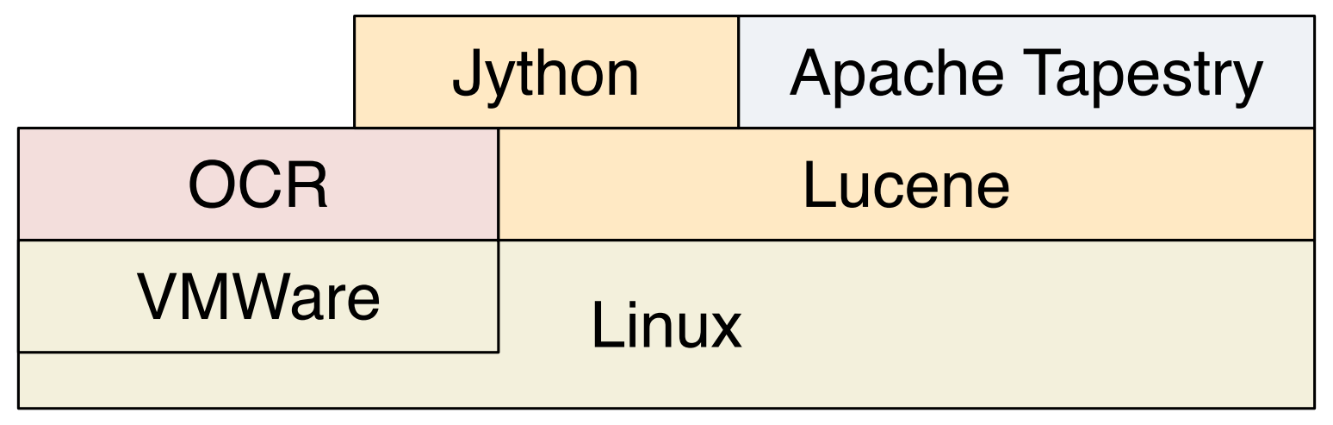
Precision, "apoptosis"

Recall, "apoptosis"

F-Rate, "apoptosis"

Precision, Recall, and F-Rate for 8 different techniques on images with captions containing "survival" and "apoptosis". Scores are for words that are *N* or more characters in length. Corpus correction is less accurate on smaller words.

Technology



Conclusion

There are several pre- and post-processing techniques that improve OCR-based text extraction. Combinations of image layout analysis (Lienhart and Wernicke, 2002; Wu et al., 1999) and context-based correction (Kukich, 1992; Ringlstetter et al., 2007) are most beneficial. Our high recall option provides an excellent basis for text indexing and search, while our high precision option works well for more general image text extraction.

References

M. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M. Wooldridge, and J. Ye. 2007. Biotext search engine: beyond abstract search. Bioinformatics, 23(16):2196–2197.

K. Kukich. 1992. Technique for automatically correcting words in text. ACM Computing Surveys, 24(4): 377–439.

R. Lienhart and A. Wernicke. 2002. Localizing and segmenting text in images and videos. IEEE Transactions on Circuits and Systems for Video Technology, 12(4):256–268, Apr.

C. Ringlstetter, K. Schulz, and S. Mihov. 2007. Adaptive text correction with web-crawled domain-dependent dictionaries. ACM Transactions on Speech and Language Processing, 4(4):1–36.

V. Wu, R. Manmatha, and E.M. Riseman. 1999. Textfinder: an automatic system to detect and recognize text in images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(11):1224–1229, Nov.

S. Xu, J. McCusker, and M. Krauthammer. 2008. Yale Image finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics, Advance Access July 9, 2008.