Automatic Highlighting of Bioscience Literature

Hudong Wang, Shannon Bradshaw, Marc Light

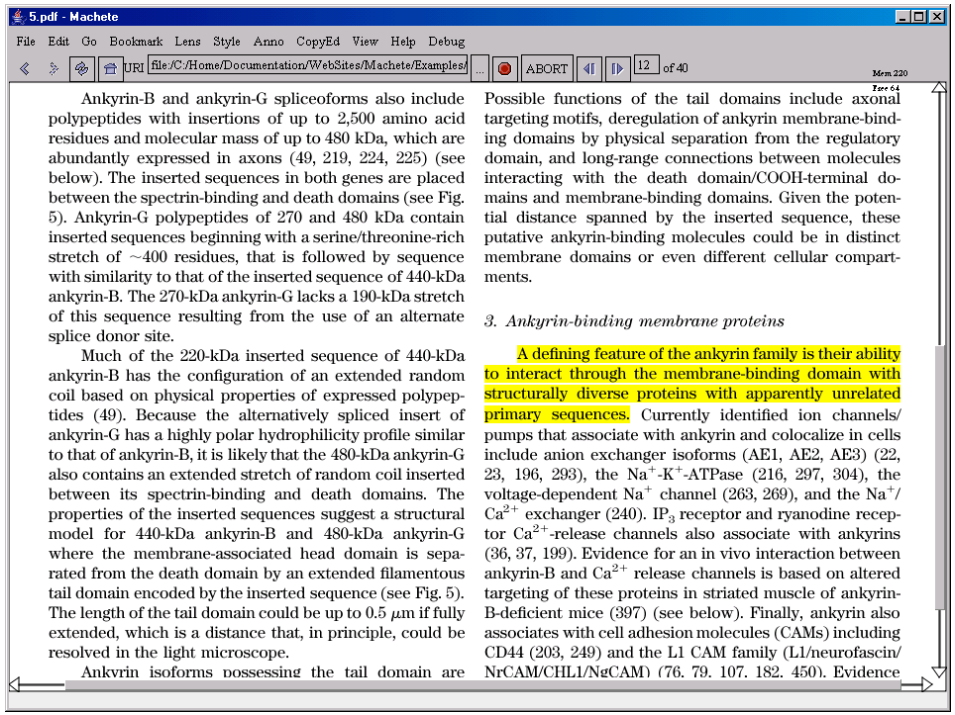
Computer Science Department
Department of Management Sciences
Linguistics Department
School of Library and Information Science

University of Iowa

Motivation

- *Bioscience researchers often read with* ***specific*** *information needs in mind*
- *Articles are often largely* ***tangential*** *to their needs*
- *Researchers* ***highlight*** *or otherwise mark relevant passages*
- ***Hypothesis****: researchers could more* ***efficiently*** *satisfy their information needs if an application could automatically highlight relevant passages*

What do we mean by highlights?



5.pdf - Machete

File Edit Go Bookmark Lens Style Anno CopyEd View Help Debug

URI file:/C:/Home/Documentation/WebSites/Machete/Examples/ ... ABORT 12 of 40

Ankyrin-B and ankyrin-G spliceoforms also include polypeptides with insertions of up to 2,500 amino acid residues and molecular mass of up to 480 kDa, which are abundantly expressed in axons (49, 219, 224, 225) (see below). The inserted sequences in both genes are placed between the spectrin-binding and death domains (see Fig. 5). Ankyrin-G polypeptides of 270 and 480 kDa contain inserted sequences beginning with a serine/threonine-rich stretch of ~400 residues, that is followed by sequence with similarity to that of the inserted sequence of 440-kDa ankyrin-B. The 270-kDa ankyrin-G lacks a 190-kDa stretch of this sequence resulting from the use of an alternate splice donor site.

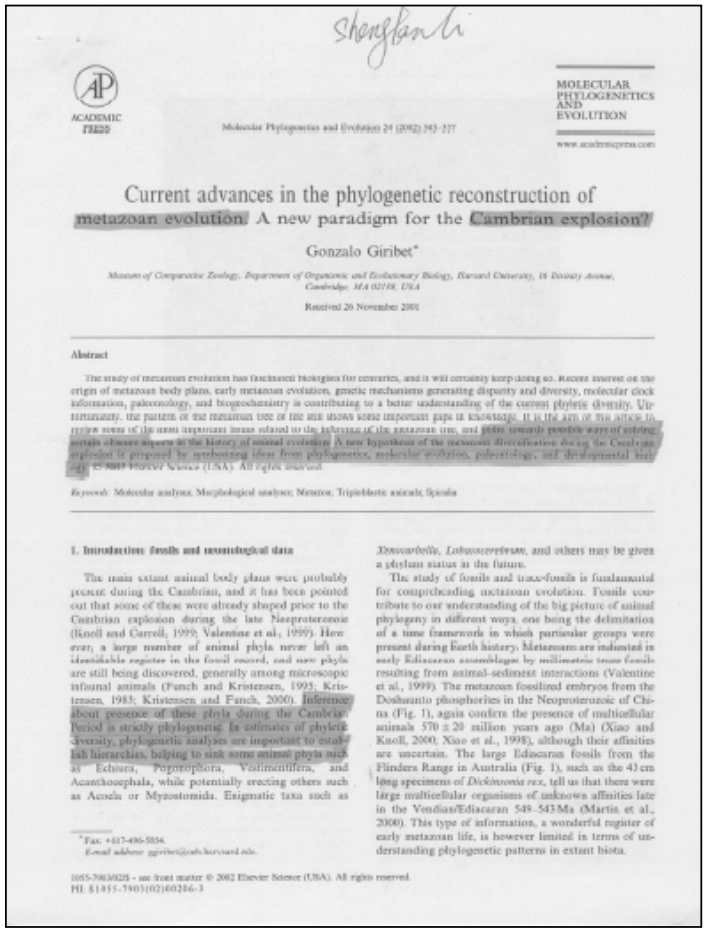
Much of the 220-kDa inserted sequence of 440-kDa ankyrin-B has the configuration of an extended random coil based on physical properties of expressed polypeptides (49). Because the alternatively spliced insert of ankyrin-G has a highly polar hydrophilicity profile similar to that of ankyrin-B, it is likely that the 480-kDa ankyrin-G also contains an extended stretch of random coil inserted between its spectrin-binding and death domains. The properties of the inserted sequences suggest a structural model for 440-kDa ankyrin-B and 480-kDa ankyrin-G where the membrane-associated head domain is separated from the death domain by an extended filamentous tail domain encoded by the inserted sequence (see Fig. 5). The length of the tail domain could be up to 0.5 μm if fully extended, which is a distance that, in principle, could be resolved in the light microscope.

Ankyrin isoforms possessing the tail domain are

Possible functions of the tail domains include axonal targeting motifs, deregulation of ankyrin membrane-binding domains by physical separation from the regulatory domain, and long-range connections between molecules interacting with the death domain/COOH-terminal domains and membrane-binding domains. Given the potential distance spanned by the inserted sequence, these putative ankyrin-binding molecules could be in distinct membrane domains or even different cellular compartments.

3. Ankyrin-binding membrane proteins

A defining feature of the ankyrin family is their ability to interact through the membrane-binding domain with structurally diverse proteins with apparently unrelated primary sequences. Currently identified ion channels/pumps that associate with ankyrin and colocalize in cells include anion exchanger isoforms (AE1, AE2, AE3) (22, 23, 196, 293), the Na^+-K^+-ATPase (216, 297, 304), the voltage-dependent Na^+ channel (263, 269), and the Na^+/Ca^{2+} exchanger (240). IP_3 receptor and ryanodine receptor Ca^{2+}-release channels also associate with ankyrins (36, 37, 199). Evidence for an in vivo interaction between ankyrin-B and Ca^{2+} release channels is based on altered targeting of these proteins in striated muscle of ankyrin-B-deficient mice (397) (see below). Finally, ankyrin also associates with cell adhesion molecules (CAMs) including CD44 (203, 249) and the L1 CAM family (L1/neurofascin/NrCAM/CHL1/NgCAM) (76, 79, 107, 182, 450). Evidence



shengfanli

ACADEMIC PRESS

Molecular Phylogenetics and Evolution 24 (2002) 345–377

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.academicpress.com

Current advances in the phylogenetic reconstruction of metazoan evolution. A new paradigm for the Cambrian explosion?

Gonzalo Giribet*

Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, 26 Divinity Avenue, Cambridge, MA 02138, USA

Received 20 November 2001

Abstract

The study of metazoan evolution has fascinated biologists for centuries, and it will certainly keep doing so. Recent interest on the origin of metazoan body plans, early metazoan evolution, genetic mechanisms generating disparity and diversity, molecular clock information, paleontology, and biogeochemistry is contributing to a better understanding of the current phyletic diversity. Unfortunately, the pattern of the metazoan tree of life still shows some important gaps in knowledge. It is the aim of this article to review some of the most important issues related to the history of the metazoan tree, and point towards possible ways of solving certain obscure aspects in the history of animal evolution. A new hypothesis of the metazoan diversification during the Cambrian explosion is proposed by synthesizing ideas from phylogenetics, molecular evolution, paleontology, and developmental biology. © 2002 Elsevier Science (USA). All rights reserved.

Keywords: Molecular analyses; Morphological analyses; Metazoa; Triploblastic animals; Spiralia

1. Introduction: fossils and neontological data

The main extant animal body plans were probably present during the Cambrian, and it has been pointed out that some of these were already shaped prior to the Cambrian explosion during the late Neoproterozoic (Knoll and Carroll, 1999; Valentine et al., 1999). However, a large number of animal phyla never left an identifiable register in the fossil record, and new phyla are still being discovered, generally among microscopic infaunal animals (Funch and Kristensen, 1995; Kristensen, 1983; Kristensen and Funch, 2000). Inference about presence of these phyla during the Cambrian Period is strictly phylogenetic. In estimates of phyletic diversity, phylogenetic analyses are important to establish hierarchies, helping to sink some animal phyla such as Echiura, Pogonophora, Vestimentifera, and Acanthocephala, while potentially erecting others such as Acoela or Myzostomida. Enigmatic taxa such as *Xenoturbella*, *Lobatocerebrum*, and others may be given a phylum status in the future.

The study of fossils and trace-fossils is fundamental for comprehending metazoan evolution. Fossils contribute to our understanding of the big picture of animal phylogeny in different ways, one being the delimitation of a time framework in which particular groups were present during Earth history. Metazoans are indicated in early Ediacaran assemblages by millimetric trace fossils resulting from animal-sediment interactions (Valentine et al., 1999). The metazoan fossilized embryos from the Doushantuo phosphorites in the Neoproterozoic of China (Fig. 1), again confirm the presence of multicellular animals 570 ± 20 million years ago (Ma) (Xiao and Knoll, 2000; Xiao et al., 1998), although their affinities are uncertain. The large Ediacaran fossils from the Flinders Range in Australia (Fig. 1), such as the 43 cm long specimens of *Dickinsonia rex*, tell us that there were large multicellular organisms of unknown affinities late in the Vendian/Ediacaran 549–543 Ma (Martin et al., 2000). This type of information, a wonderful register of early metazoan life, is however limited in terms of understanding phylogenetic patterns in extant biota.

* Fax: +617-496-5854.
E-mail address: ggiribet@oeb.harvard.edu.

1055-7903/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
PII: S1055-7903(02)00286-3

Goal: automate a first cut at highlighting to enable a quick scan of the paper

Our Data

- *13 articles that **had already been** highlighted*
  - *a mix of topics, a mix of computer- and paper-based highlighting, 2 biologists*
- *Asked what their information need **had been***
  - *evolution, coevolution "RNA worrld", retroelement, "retroelement ancestor hypothesis", mobility, mobile*
- *http://que.info-science.uiowa.edu/~light/research/data/lightBioLink2005HighlightingData.tgz*

#highlighted sents

More on the Data

	Relev.	Tot.	Article	Annot.	Qlen	1stHlen
	24	253	Collins	D	9	15
	41	957	Lamb	D	14	45
	22	721	Lohne	D	15	18
	9	356	Toor	D	10	23
	13	483	Wiens	D	6	58
	21	187	Ferl	T	4	26
	13	358	Hilt	T	4	14
	8	568	Mori	T	4	21
	18	279	Nakrie	T	4	32
	13	258	Roberts	T	4	3
	9	749	Sehnke	T	4	2
	28	175	SHenry	T	4	21
	13	513	Smith	T	4	25
mean	*18*	*450*			*10*	*23*

How is this Different From Normal Document Search?

- ***Not a needle in a haystack****: fewer sentences in an article than documents in a collection*
- ***Not many words to work with****: passages have fewer words than documents*
  - *query expansion may be crucial*
- ***"Relevant"*** *may mean something different*

Automatic Highlighting System

- *Basic "search" engine and treated each sentence as a document*
- *Two types of queries:*
  - *the keywords provided*
  - *the first highlighted passage*
- *Expanded the query based on definitions*
- *Sets of definitions culled from the web*

Questions We Asked

- *How well does a standard retrieval work (okapi)?*
- *What is a better query?*
  - *keywords*
  - *example passage*
- *Does query expansion based on definitions help?*
- *Do multiple definitions help more?*

Web Definition Sets

Google Search: define:apoptosis

http://www.google.com/search?client=safari&rls=en&q=define:apoptosis&ie=UTF-8&oe

define:apoptosis

iHawk GoogleScho PgresDoc javaDoc psql PubMed forums home FileComplaint cvsDoc ArtReq domin macheteWiki Apple

The Machete Project | PBR: The Protein Browse... | Google Search: define:ap...

Google

Web Images Groups News Froogle Local New! more »

define:apoptosis Search Advanced Search Preferences

Web

Definitions of **apoptosis** on the Web:

- "Cellular suicide," also known as programmed cell death. HIV may induce apoptosis in both infected and uninfected immune system cells. Normally when CD4+ T cells mature in the thymus gland, a small proportion of these cells are unable to distinguish self from nonself. Because these cells would otherwise attack the body's own tissues, they receive a biochemical signal from other cells that results in apoptosis. See Tumor Necrosis Factor.
  www.thebody.com/hivatis/glossary/a.html

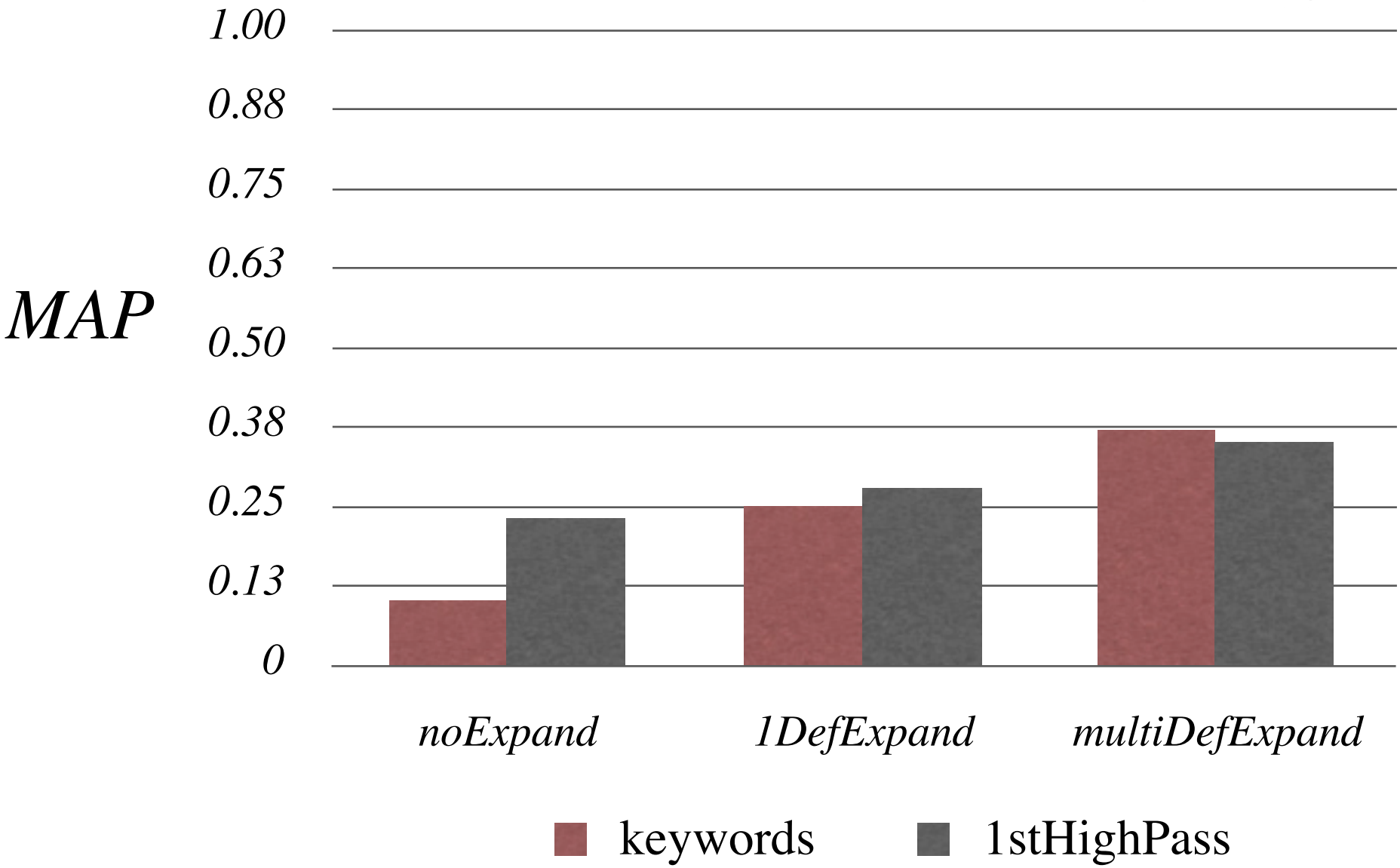
- Programmed cell death.
  www.usc.edu/hsc/dental/opath/Chapters/DictionaryA.html

- programmed cell death, or "cell suicide"; a form of cell death in which a controlled sequence of events (or program) leads to the elimination of cells without releasing harmful substances into the surrounding area. Many types of cell damage can trigger apoptosis, and it also occurs normally during development of the nervous system and other parts of the body. Strictly speaking, the term apoptosis refers only to the structural changes cells go through, and programmed cell death refers to the complete underlying process, but the terms are often used synonymously.
  www.ninds.nih.gov/news_and_events/proceedings/sci_report_pr.htm

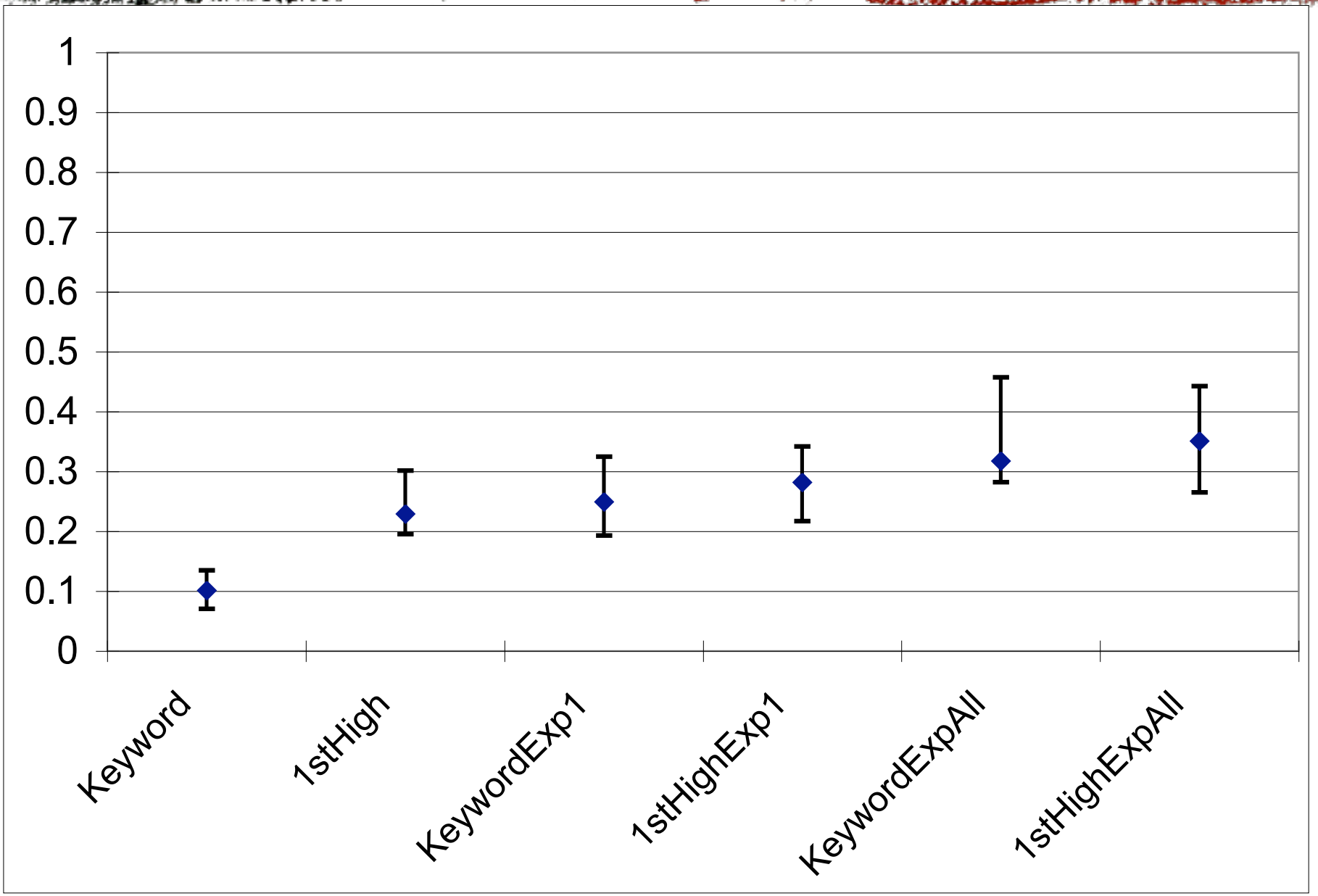
- A normal cellular process involving a genetically programmed series of events leading to the death of a cell.
  www.nexxusscotland.com/media_centre/glossary.html

- a form of cell death involving shrinking of the cell and eventual disposal of the internal elements of the cell by the body's immune system. Apoptosis is an active, non-toxic form of cell suicide that does not induce an inflammatory response. It is

Results


MAP
1.00
0.88
0.75
0.63
0.50
0.38
0.25
0.13
0
noExpand
1DefExpand
multiDefExpand
keywords
1stHighPass

Results
Bootstrapped Confidence Intervals



Questions "Answered"

- ***Standard retrieval****: 0.23 MAP (not good)*
- ***Example highlighted regions*** *are better than keywords specifying information need*
- ***Definition-based*** *query expansion helps*
- ***Multiple*** *definitions helps more*

- ***Contributions****: introduced task, baseline results, multi-definition-based query expansion*

More Data To Come

- *Graduate seminar on evolutionary biology*
- *7 students all marked up the* ***same*** *16 articles*
- *All hardcopy markup*
- *We have scanned the hardcopy markup*
- *We have ASCII of the articles and are laboriously creating corresponding XML markup (using Callisto (thanks MITRE))*

Future Work On Highlighting

- *Create application and field it*
- *Compare and contrast other query expansion methods*
- *Find and use collocations in definitions "cell death" instead of "cell" and "death"*