UNIVERSITAS TARTUENSIS
1632

Erik Jaaniso

Automatic mapping of free texts to bioinformatics ontology terms

Master's Thesis

Supervisor: Hedi Peterson, PhD

06.06.2016

Motivation

Thousands of tools and services in bioinformatics

Motivation

Thousands of tools and services in bioinformatics

⇒ how to find & not make a duplicate?

Motivation

Thousands of tools and services in bioinformatics

⇒ how to find & not make a duplicate?

💡 Collect metadata in a database

Motivation

Thousands of tools and services in bioinformatics

⇒ how to find & not make a duplicate?

💡 Collect metadata in a database

⇒ how to find from database?

Motivation

Thousands of tools and services in bioinformatics

⇒ how to find & not make a duplicate?

💡 Collect metadata in a database

⇒ how to find from database?

💡 Categorise and annotate using a standard vocabulary

Motivation

Thousands of tools and services in bioinformatics

⇒ how to find & not make a duplicate?

💡 Collect metadata in a database

⇒ how to find from database?

💡 Categorise and annotate using a standard vocabulary

⇒ done manually by a *curator*

Motivation

Thousands of tools and services in bioinformatics
⇒ how to find & not make a duplicate?
💡 Collect metadata in a database
⇒ how to find from database?
💡 Categorise and annotate using a standard vocabulary
⇒ done manually by a *curator*

Mapping tools & services to vocabulary terms requires:
☞ Motivation

Motivation

Thousands of tools and services in bioinformatics

⇒ how to find & not make a duplicate?

💡 Collect metadata in a database

⇒ how to find from database?

💡 Categorise and annotate using a standard vocabulary

⇒ done manually by a *curator*

Mapping tools & services to vocabulary terms requires:

☞ Motivation

☞ Time

Motivation

Thousands of tools and services in bioinformatics
⇒ how to find & not make a duplicate?
💡 Collect metadata in a database
⇒ how to find from database?
💡 Categorise and annotate using a standard vocabulary
⇒ done manually by a *curator*

Mapping tools & services to vocabulary terms requires:
☞ Motivation
☞ Time
☞ Knowledge (of both tool/service and vocabulary)

Motivation

Thousands of tools and services in bioinformatics

⇒ how to find & not make a duplicate?

💡 Collect metadata in a database

⇒ how to find from database?

💡 Categorise and annotate using a standard vocabulary

⇒ done manually by a *curator*

Mapping tools & services to vocabulary terms requires:

☞ Motivation

☞ Time

☞ Knowledge (of both tool/service and vocabulary)

The curator might lack any of them.

Motivation

Thousands of tools and services in bioinformatics

⇒ how to find & not make a duplicate?

💡 Collect metadata in a database

⇒ how to find from database?

💡 Categorise and annotate using a standard vocabulary

⇒ done manually by a *curator*

Mapping tools & services to vocabulary terms requires:

☞ Motivation

☞ Time

☞ Knowledge (of both tool/service and vocabulary)

The curator might lack any of them. Can we help him?

Automatic mapping of free texts to bioinformatics **ontology terms**

Automatic mapping of free texts to bioinformatics **ontology terms**

Ontology (philosophy) – what “things” exist

- ⇒ “things” represented by *concepts*, labelled by *terms*
- ⇒ use different terms for same concept, or *vice versa*

Automatic mapping of free texts to bioinformatics **ontology terms**

Ontology (philosophy) – what “things” exist

⇒ “things” represented by *concepts*, labelled by *terms*

⇒ use different terms for same concept, or *vice versa*

Humans good at disambiguating and guessing meaning

⇒ but computers not

Automatic mapping of free texts to bioinformatics **ontology terms**

Ontology (philosophy) – what “things” exist

⇒ “things” represented by *concepts*, labelled by *terms*

⇒ use different terms for same concept, or *vice versa*

Humans good at disambiguating and guessing meaning

⇒ but computers not

Ontology (CS) – make knowledge computationally useful

Automatic mapping of free texts to bioinformatics **ontology terms**

Ontology (philosophy) – what “things” exist

⇒ “things” represented by *concepts*, labelled by *terms*

⇒ use different terms for same concept, or *vice versa*

Humans good at disambiguating and guessing meaning

⇒ but computers not

Ontology (CS) – make knowledge computationally useful

☞ conceptualise things into classes (e.g., chair)

☞ describe relationships (e.g., *is a* furniture)

Automatic mapping of free texts to bioinformatics **ontology terms**

Ontology (philosophy) – what “things” exist

⇒ “things” represented by *concepts*, labelled by *terms*

⇒ use different terms for same concept, or *vice versa*

Humans good at disambiguating and guessing meaning

⇒ but computers not

Ontology (CS) – make knowledge computationally useful

☞ conceptualise things into classes (e.g., chair)

☞ describe relationships (e.g., *is a* furniture)

To better query, browse and share knowledge in a domain

Automatic mapping of free texts to **bioinformatics ontology** terms

Automatic mapping of free texts to **bioinformatics ontology** terms

EDAM – simple bioinformatics ontology (3218 concepts)

Automatic mapping of free texts to **bioinformatics ontology** terms

EDAM – simple bioinformatics ontology (3218 concepts)

4 main sub-ontologies or *branches*:

- ☞ **topic** – “Data visualisation”, “Proteomics”
- ☞ **operation** – “Visualisation”, “Sequence alignment”
- ☞ **data** – “Image”, “Sequence”
- ☞ **format** – “PNG”, “FASTA”

Automatic mapping of free texts to **bioinformatics ontology** terms

EDAM – simple bioinformatics ontology (3218 concepts)

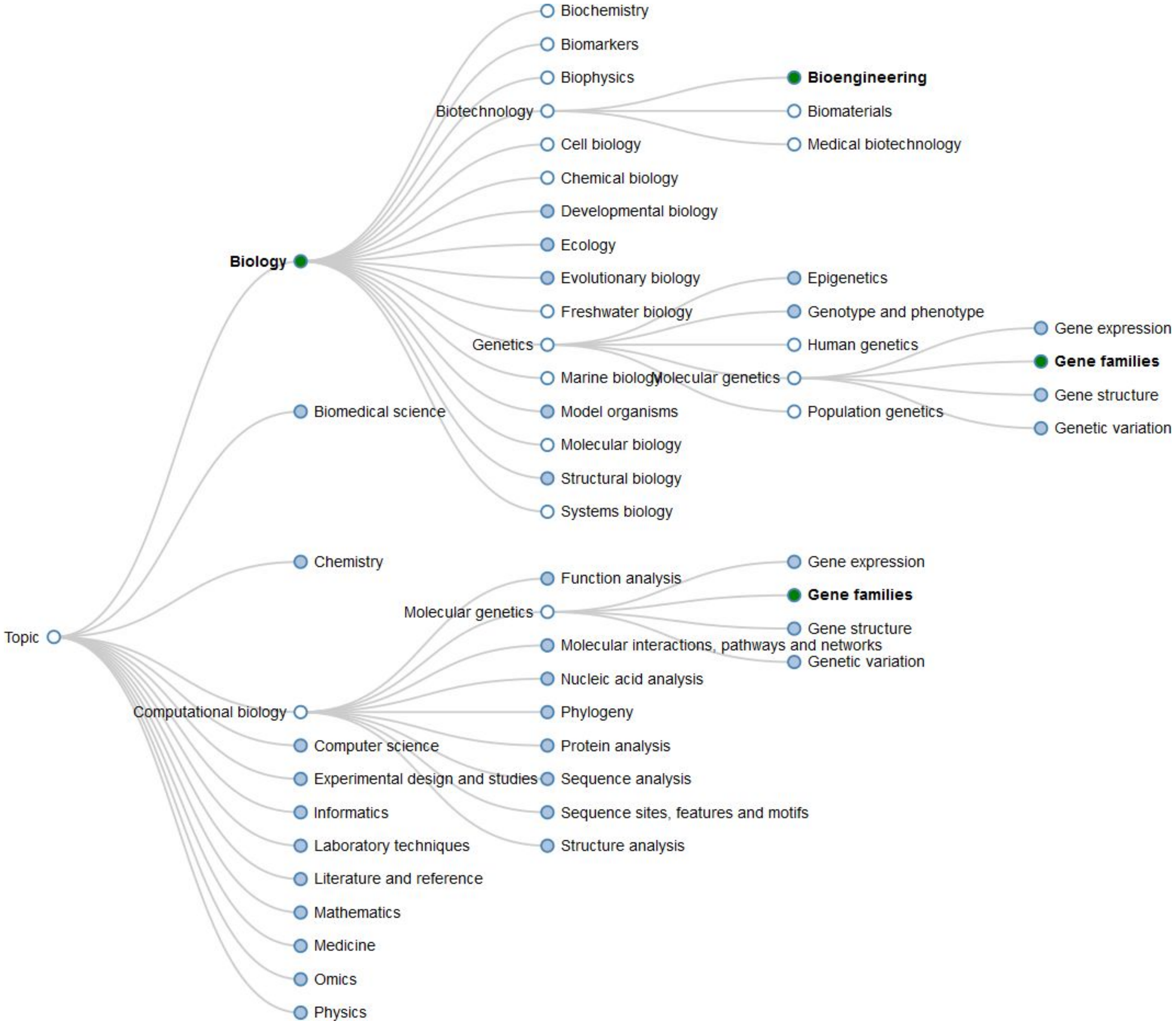
4 main sub-ontologies or *branches*:

- ☞ **topic** – “Data visualisation”, “Proteomics”
- ☞ **operation** – “Visualisation”, “Sequence alignment”
- ☞ **data** – “Image”, “Sequence”
- ☞ **format** – “PNG”, “FASTA”

Concept ID is URI and it has *parts*:

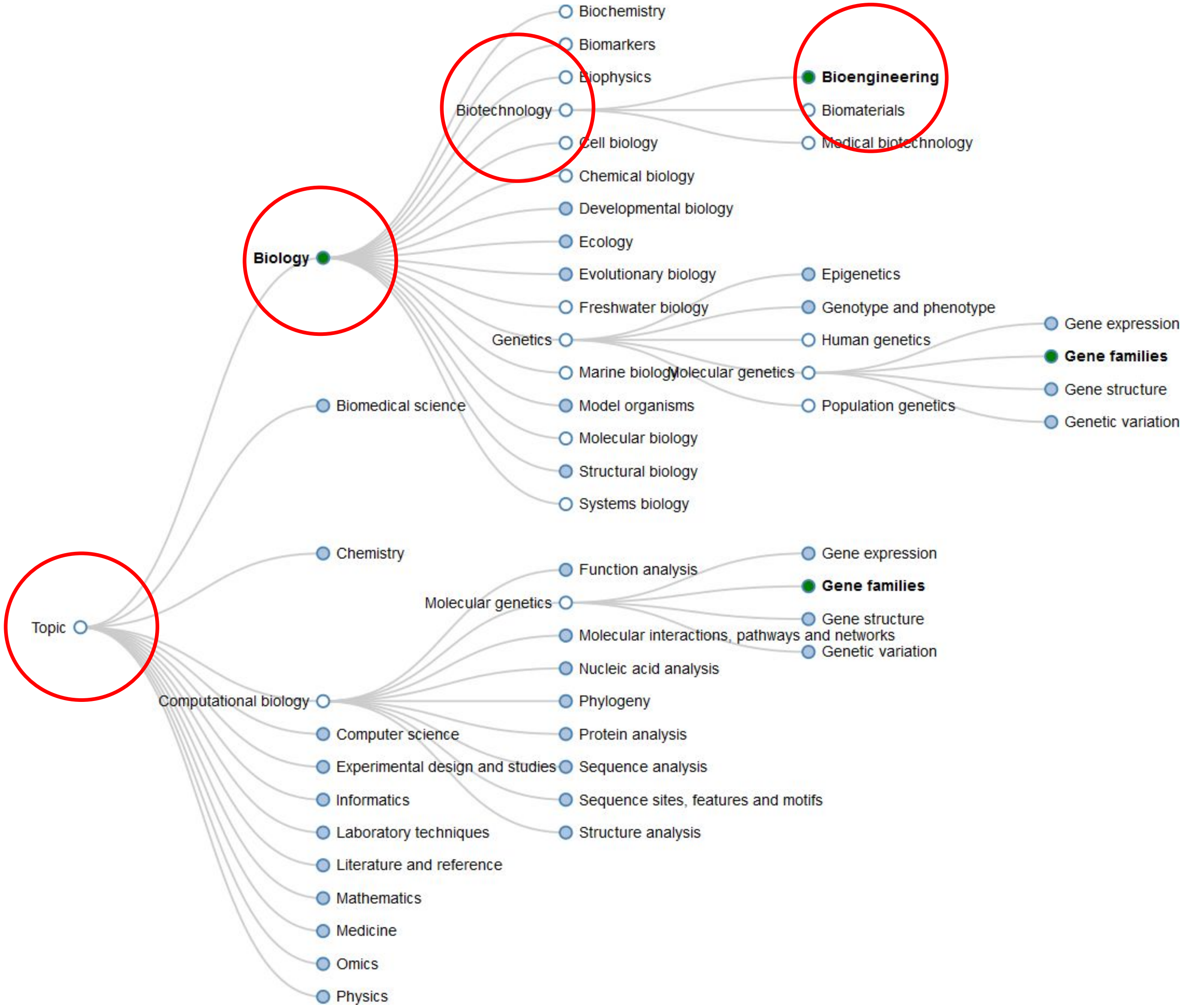
- ☞ **label**/term – “Visualisation”
- ☞ **synonyms** (exact/narrow/broad) – “Rendering”
- ☞ **definition** & comment – “Visualise, plot or render (graphically) biomolecular data such as molecular sequences or structures.”

Automatic mapping of free texts to **bioinformatics ontology** terms



http://rainbio.france-bioinformatique.fr/rainbio/browseEdam

Automatic mapping of free texts to **bioinformatics ontology** terms



http://rainbio.france-bioinformatique.fr/rainbio/browseEdam

Automatic mapping of **free texts** to bioinformatics ontology terms

Automatic mapping of **free texts** to bioinformatics ontology terms

Main tools & services metadata database used:

bio.tools – ELIXIR Tools and Data Services Registry

⇒ as of writing thesis, 2402 entries

Automatic mapping of **free texts** to bioinformatics ontology terms

Main tools & services metadata database used:
bio.tools – ELIXIR Tools and Data Services Registry
⇒ as of writing thesis, 2402 entries

An entry is a collection of free texts (*query*), with *parts*:

Automatic mapping of **free texts** to bioinformatics ontology terms

Main tools & services metadata database used:
bio.tools – ELIXIR Tools and Data Services Registry
⇒ as of writing thesis, 2402 entries

An entry is a collection of free texts (*query*), with *parts*:

☞ **name** – “WEBnma”

☞ **description** – “provides users with quick, automated computation and analysis of low-frequency normal modes for protein structures.”

☞ **publication** – 10.1186/s12859-014-0427-6, 10.1186/1471-2105-6-52

☞ **homepage** – http://apps.cbu.uib.no/webnma

☞ **documentation** – http://apps.cbu.uib.no/webnma/howto

Automatic mapping of free texts to bioinformatics ontology terms

Automatic mapping of free texts to bioinformatics ontology terms

Match words between tool/service parts and concept parts

⇒ if matches found, suggest concept as annotation

Automatic mapping of free texts to bioinformatics ontology terms

Match words between tool/service parts and concept parts

⇒ if matches found, suggest concept as annotation

"... automated computation and analysis of low-frequency normal modes for protein structures."

"Protein flexibility and motion analysis"

Automatic mapping of free texts to bioinformatics ontology terms

Match words between tool/service parts and concept parts
⇒ if matches found, suggest concept as annotation

"... automated computation and analysis of low-frequency normal modes for protein structures."	"Protein flexibility and motion analysis"
"... automated computation and analysis of low-frequency normal modes for protein structures."	"Protein structure analysis"

Automatic mapping of free texts to bioinformatics ontology terms

Match words between tool/service parts and concept parts
⇒ if matches found, suggest concept as annotation

"… automated computation and analysis of low-frequency normal modes for protein structures."	"Protein flexibility and motion analysis"
"… automated computation and analysis of low-frequency normal modes for protein structures."	"Protein structure analysis"

But, how to order the suggestions?

Automatic mapping of free texts to bioinformatics ontology terms

Match words between tool/service parts and concept parts
⇒ if matches found, suggest concept as annotation

"… automated computation and analysis of low-frequency normal modes for protein structures."	"Protein flexibility and motion analysis"
"… automated computation and analysis of low-frequency normal modes for protein structures."	"Protein structure analysis"

But, how to order the suggestions?
☞ more words matching

Automatic mapping of free texts to bioinformatics ontology terms

Match words between tool/service parts and concept parts
⇒ if matches found, suggest concept as annotation

"… automated computation and analysis of low-frequency normal modes for protein structures."	"Protein flexibility and motion analysis"
"… automated computation and analysis of low-frequency normal modes for protein structures."	"Protein structure analysis"

But, how to order the suggestions?

☞ more words matching

☞ more context preserved

Automatic mapping of free texts to bioinformatics ontology terms

Match words between tool/service parts and concept parts
⇒ if matches found, suggest concept as annotation

"… automated computation and analysis of low-frequency normal modes for protein structures."	"Protein flexibility and motion analysis"
"… automated computation and analysis of low-frequency normal modes for protein structures."	"Protein structure analysis"

But, how to order the suggestions?

- ☞ more words matching
- ☞ more context preserved
- ☞ more parts matched

Automatic mapping of free texts to bioinformatics ontology terms

Match words between tool/service parts and concept parts
⇒ if matches found, suggest concept as annotation

"... automated computation and analysis of low-frequency normal modes for protein structures."	"Protein flexibility and motion analysis"
"... automated computation and analysis of low-frequency normal modes for protein structures."	"Protein structure analysis"

But, how to order the suggestions?

☞ more words matching
☞ more context preserved
☞ more parts matched
☞ more "important" words matching

Fetching publication content

Fetching publication content

10.1186/s12859-014-0427-6

Fetching publication content

10.1186/s12859-014-0427-6

⇓

Title: “WEBnm@ v2.0: Web server and services for comparing protein flexi...”

Keywords: “Elastic network models”, “Normal mode analysis”, “Web-tool”, ...

Abstract: “Normal mode analysis (NMA) using elastic network models is a reliable and cost-effective computational method to characterise ...”

Fulltext: “Protein dynamics is defined as the time-dependent changes in the structure of a protein, which includes equilibrium fluctuations governing ...”

Fetching publication content

10.1186/s12859-014-0427-6

⇓

Title: “WEBnm@ v2.0: Web server and services for comparing protein flexi...”
Keywords: “Elastic network models”, “Normal mode analysis”, “Web-tool”, ...
Abstract: “Normal mode analysis (NMA) using elastic network models is a reliable and cost-effective computational method to characterise ...”
Fulltext: “Protein dynamics is defined as the time-dependent changes in the structure of a protein, which includes equilibrium fluctuations governing ...”

☞ Query publication databases (like PubMed) by ID

Fetching publication content

10.1186/s12859-014-0427-6

⇓

Title: “WEBnm@ v2.0: Web server and services for comparing protein flexi...”
Keywords: “Elastic network models”, “Normal mode analysis”, “Web-tool”, ...
Abstract: “Normal mode analysis (NMA) using elastic network models is a reliable and cost-effective computational method to characterise ...”
Fulltext: “Protein dynamics is defined as the time-dependent changes in the structure of a protein, which includes equilibrium fluctuations governing ...”

- ☞ Query publication databases (like PubMed) by ID
- ☞ Or, from publisher website (custom extraction rules)

Fetching publication content

10.1186/s12859-014-0427-6

⇓

Title: “WEBnm@ v2.0: Web server and services for comparing protein flexi...”
Keywords: “Elastic network models”, “Normal mode analysis”, “Web-tool”, ...
Abstract: “Normal mode analysis (NMA) using elastic network models is a reliable and cost-effective computational method to characterise ...”
Fulltext: “Protein dynamics is defined as the time-dependent changes in the structure of a protein, which includes equilibrium fluctuations governing ...”

- ☞ Query publication databases (like PubMed) by ID
- ☞ Or, from publisher website (custom extraction rules)
- ☞ Publication parts extracted from XML, HTML or PDF

Fetching homepage & docs

Fetching homepage & docs

http://apps.cbu.uib.no/webnma

Fetching homepage & docs

http://apps.cbu.uib.no/webnma

⇓

“About WEBnm@ provides users with quick, automated computation and analysis of low-frequency normal modes for protein structures. The computation performed through our server should help the user understand whether a given protein can undergo large amplitude movements, and …”

http://apps.cbu.uib.no/webnma/howto

⇓

“HowTo Single Analysis Comparative Analysis Examples Single Analysis Comparative Analysis Other input examples Single Analysis Submit a structure file in the pdb format and our server will calculate the lowest frequency normal modes of your molecule. You will then be offered different types of …”

Pre-processing

“WEBnm@ provides users with quick, automated computation and analysis of low-frequency normal modes for protein structures.”

Pre-processing

“WEBnm@ provides users with quick, automated computation and analysis of low-frequency normal modes for protein structures.”

⇓

Punctuation removal and tokenisation

[webnm, provides, users, with, quick, automated, computation, and, analysis, of, lowfrequency, normal, modes, for, protein, structures]

⇓

Stop words removal

[webnm, provides, users, quick, automated, computation, analysis, lowfrequency, normal, modes, protein, structures]

⇓

Stemming

[webnm, provid, user, quick, autom, comput, analysi, lowfrequ, normal, mode, protein, structur]

Mapping algorithm

Mapping algorithm

[protein, structur, analysi] ⇒ [webnm, provid, user, quick, autom, comput, analysi, lowfrequ, normal, mode, protein, structur]

↳ **Score**

Mapping algorithm

[protein, structur, analysi] ⇒ [webnm, provid, user, quick, autom, comput,
analysi, lowfrequ, normal, mode, protein, structur]

↳ **Score**

For each concept a match score, so annotations are ranked

Mapping algorithm

[protein, structur, analysi] ⇒ [webnm, provid, user, quick, autom, comput,
analysi, lowfrequ, normal, mode, protein, structur]

↳ **Score**

For each concept a match score, so annotations are ranked

Performance → benchmark against manual annotations

Mapping algorithm

[protein, structur, analysi] ⇒ [webnm, provid, user, quick, autom, comput, analysi, lowfrequ, normal, mode, protein, structur]

↳ **Score**

For each concept a match score, so annotations are ranked

Performance → benchmark against manual annotations

☞ mean recall (how much found among top *n*)

☞ mean average precision (how are ranked)

Mapping algorithm

[protein, structur, analysi] ⇒ [webnm, provid, user, quick, autom, comput, analysi, lowfrequ, normal, mode, protein, structur]

↳ **Score**

For each concept a match score, so annotations are ranked

Performance → benchmark against manual annotations

☞ mean recall (how much found among top *n*)

☞ mean average precision (how are ranked)

Can be outdated, incomplete ⇒ but best we have

Mapping algorithm

[protein, structur, analysi] ⇒ [webnm, provid, user, quick, autom, comput, analysi, lowfrequ, normal, mode, protein, structur]

↳ **Score**

For each concept a match score, so annotations are ranked

Performance → benchmark against manual annotations

☞ mean recall (how much found among top *n*)

☞ mean average precision (how are ranked)

Can be outdated, incomplete ⇒ but best we have

Manually tune algorithm parameters for best performance

Mapping algorithm features

👍👎 Approximate matching

👍 Proximity matching

👍👍👎👎 Inverse document frequency

👍 Non-linear scaling of match count

👍👍 Bi-directional matching

👍 Combining parts’ scores (weighted average)

👍👎 Stop words removal & Stemming

“WEBnma” to “Protein flexibility and motion analysis”

“WEBnma”

Tool

WEBnma

description

[webnm, provid, user, quick, autom, comput, analysi, lowfrequ, normal, mode, protein, structur]

publication title

[webnm, v20, web, server, servic, compar, protein, flexibl]

publication abstract

[background, normal, mode, analysi, nma, us, elast, network, model, reliabl, costeffect, comput, method, characteris, protein, flexibl, extens, dynam, further, insight, dynamicsfunct, relationship, can, gain, compar, protein, motion, between, protein, homolog, function, classif, can, achiev, compar, normal, mode, obtain, from, set, evolutionari, relat, protein, result, we, have, develop, autom, tool, compar, nma, set, prealign, protein, structur, user, can, submit, sequenc, align, fasta, format, correspond, coordin, file, protein, data, bank, pdb, format, comput, normalis, squar, atom, fluctuat, atom, deform, energi, submit, structur, can, easili, compar, graph, provid, web, user, interfac, web, server, provid, pairwis, comparison, dynam, all, protein, includ, submit, set, us, two, measur, root, mean, squar, inner, product, bhattacharyya, coeffici, compar, analysi, ha, been, implement, our, web, server, nma, webnm, which, also, provid, recent, upgrad, function, nma, singl, protein, structur, includ, new, visualis, protein, motion, visualis, interresidu, correl, analysi, conform, chang, us, overlap, analysi, addit, programmat, access, webnm, now, avail, through, soapbas, web, servic, webnm, avail, http, appscbuuibno, webnma, conclus, webnm, v20, onlin, tool, offer, uniqu, capabl, compar, nma, multipl, protein, structur, along, conveni, web, interfac, power, comput, resourc, sever, method, mode, analys, webnm, facilit, assess, protein, flexibl, within, protein, famili, superfamili, analys, can, give, good, view, how, structur, move, how, flexibl, conserv, over, differ, structur]

“Protein flexibility and motion analysis”

Concept http://edamontology.org/operation_0244

label

[protein, flexibl, motion, analysi]

definition

[analys, flexibl, motion, protein, structur]

comment

[us, concept, analysi, flexibl, rigid, residu, local, chain, deform, region, undergo, conform, chang, molecular, vibrat, fluctuat, dynam, domain, motion, other, largescal, structur, transit, protein, structur]

“WEBnma” to “Protein flexibility and motion analysis”

Tool

WEBnma

description

[webnm, provid, user, quick, autom, comput, analysi, lowfrequ, normal, mode, protein, structur]

publication title

[webnm, v20, web, server, servic, compar, protein, flexibl]

publication abstract

[background, normal, mode, analysi, nma, us, elast, network, model, reliabl, costeffect, comput, method, characteris, protein, flexibl, extens, dynam, further, insight, dynamicsfunct, relationship, can, gain, compar, protein, motion, between, protein, homolog, function, classif, can, achiev, compar, normal, mode, obtain, from, set, evolutionari, relat, protein, result, we, have, develop, autom, tool, compar, nma, set, prealign, protein, structur, user, can, submit, sequenc, align, fasta, format, correspond, coordin, file, protein, data, bank, pdb, format, comput, normalis, squar, atom, fluctuat, atom, deform, energi, submit, structur, can, easili, compar, graph, provid, web, user, interfac, web, server, provid, pairwis, comparison, dynam, all, protein, includ, submit, set, us, two, measur, root, mean, squar, inner, product, bhattacharyya, coeffici, compar, analysi, ha, been, implement, our, web, server, nma, webnm, which, also, provid, recent, upgrad, function, nma, singl, protein, structur, includ, new, visualis, protein, motion, visualis, interresidu, correl, analysi, conform, chang, us, overlap, analysi, addit, programmat, access, webnm, now, avail, through, soapbas, web, servic, webnm, avail, http, appscbuuibno, webnma, conclus, webnm, v20, onlin, tool, offer, uniqu, capabl, compar, nma, multipl, protein, structur, along, conveni, web, interfac, power, comput, resourc, sever, method, mode, analys, webnm, facilit, assess, protein, flexibl, within, protein, famili, superfamili, analys, can, give, good, view, how, structur, move, how, flexibl, conserv, over, differ, structur]

⇔

Concept http://edamontology.org/operation_0244

label

[protein, flexibl, motion, analysi]

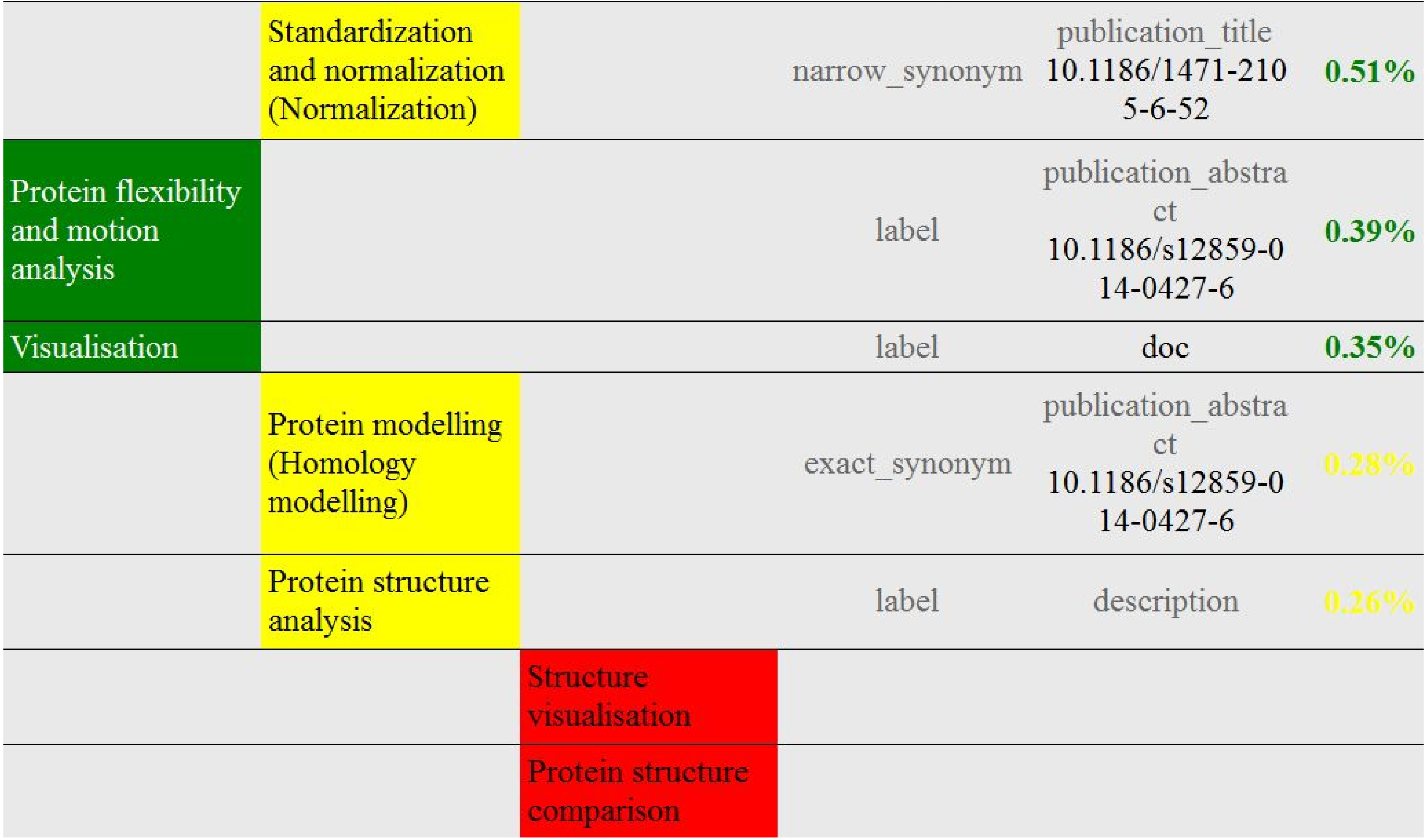
definition

[analys, flexibl, motion, protein, structur]

comment

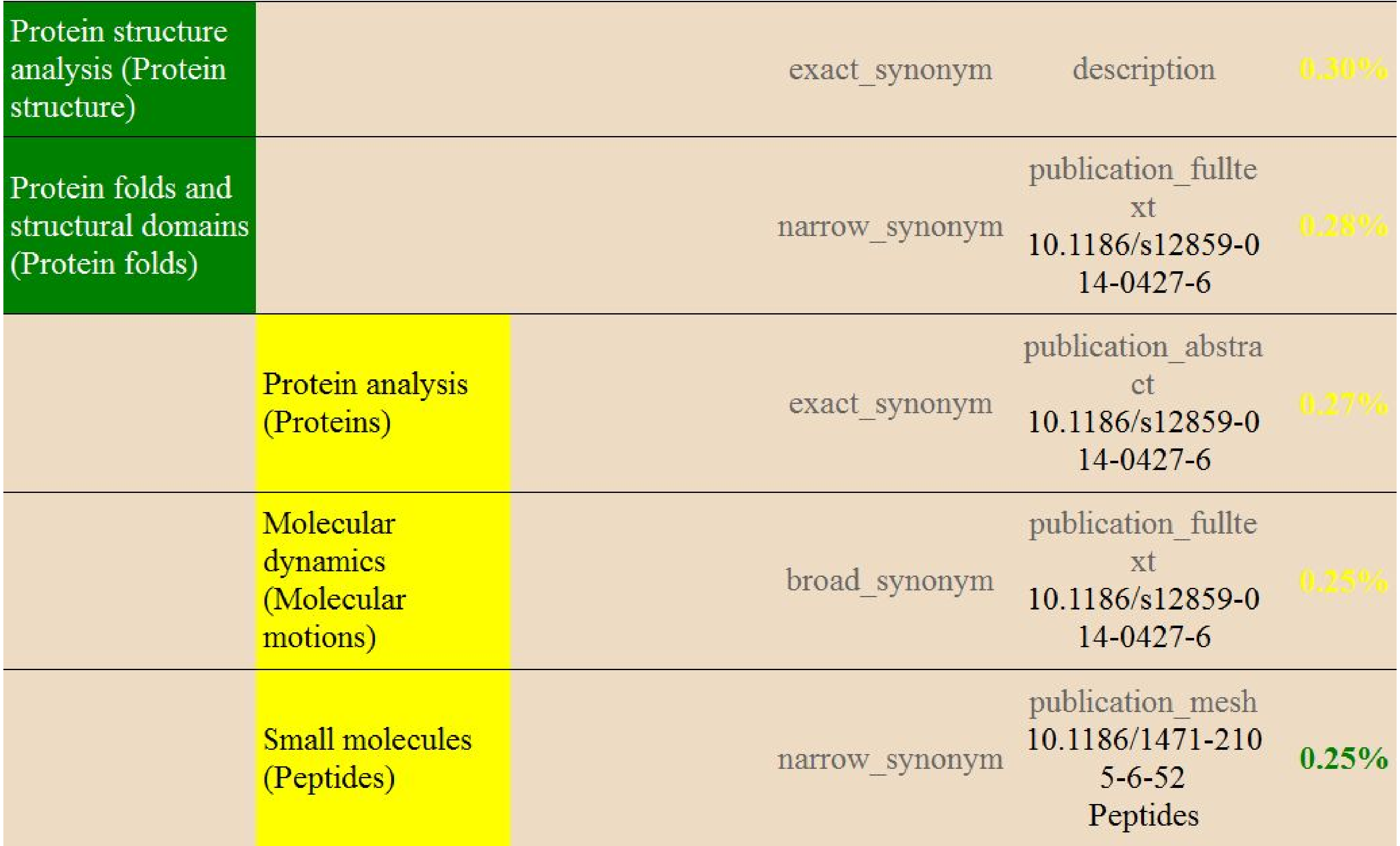
[us, concept, analysi, flexibl, rigid, residu, local, chain, deform, region, undergo, conform, chang, molecular, vibrat, fluctuat, dynam, domain, motion, other, largescal, structur, transit, protein, structur]

“WEBnma” to *operation* branch



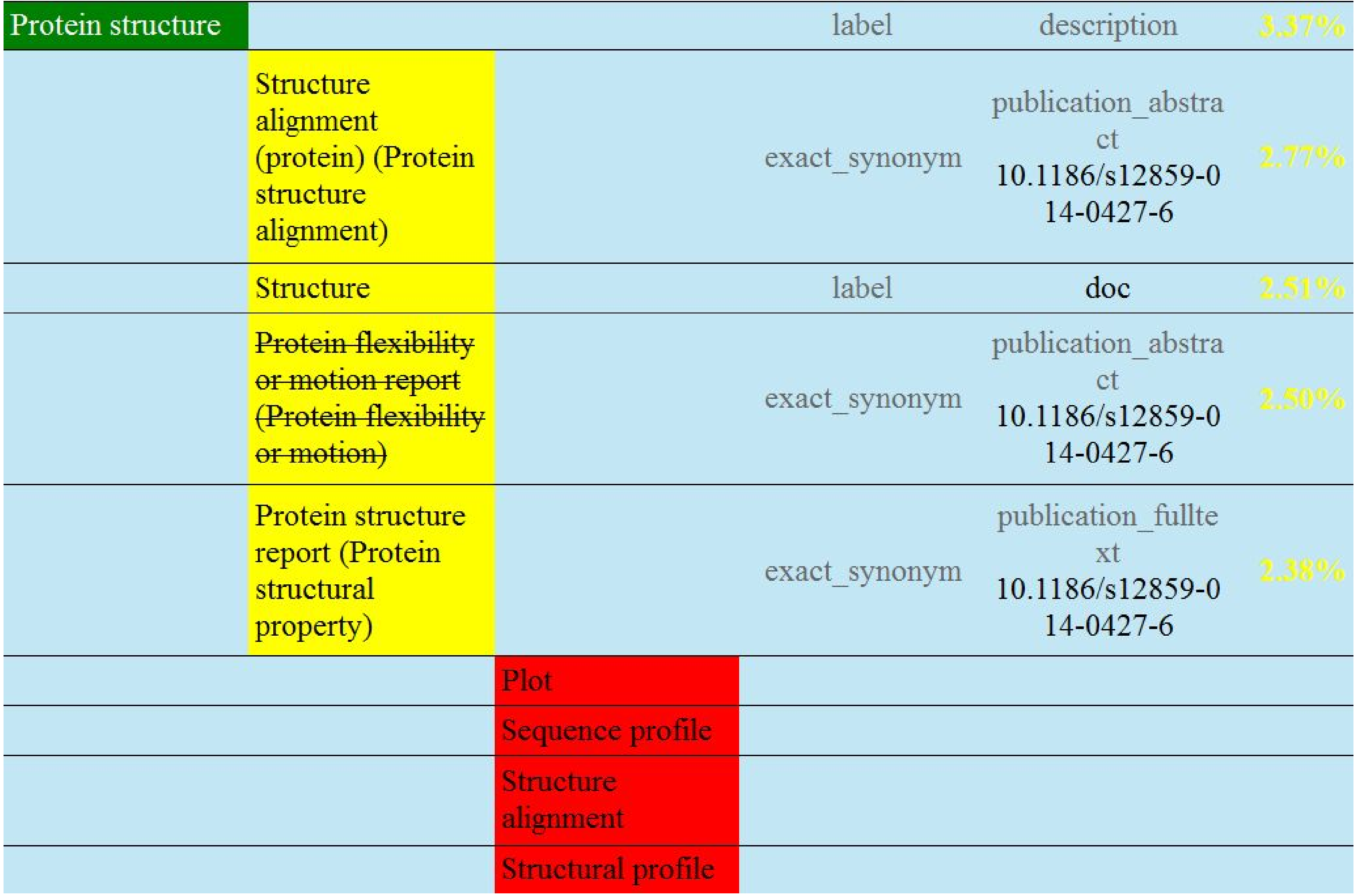
TP	FP	FN	Concept	Query	Score
	Standardization and normalization (Normalization)		narrow_synonym	publication_title 10.1186/1471-2105-6-52	0.51%
Protein flexibility and motion analysis			label	publication_abstract 10.1186/s12859-014-0427-6	0.39%
Visualisation			label	doc	0.35%
	Protein modelling (Homology modelling)		exact_synonym	publication_abstract 10.1186/s12859-014-0427-6	0.28%
	Protein structure analysis		label	description	0.26%
		Structure visualisation			
		Protein structure comparison			

“WEBnma” to *topic* branch



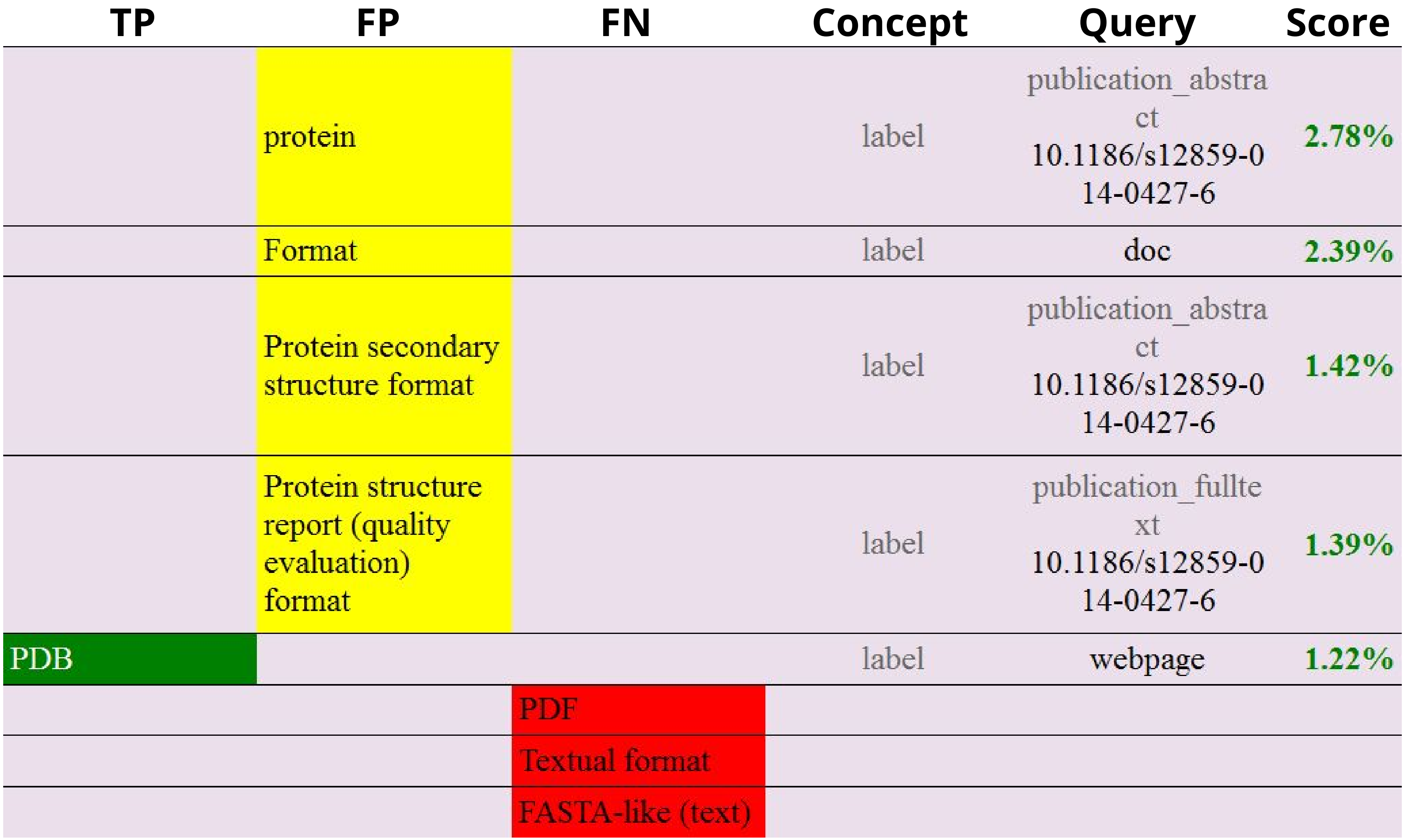
TP	FP	FN	Concept	Query	Score
Protein structure analysis (Protein structure)			exact_synonym	description	0.30%
Protein folds and structural domains (Protein folds)			narrow_synonym	publication_fulltext 10.1186/s12859-014-0427-6	0.28%
	Protein analysis (Proteins)		exact_synonym	publication_abstract 10.1186/s12859-014-0427-6	0.27%
	Molecular dynamics (Molecular motions)		broad_synonym	publication_fulltext 10.1186/s12859-014-0427-6	0.25%
	Small molecules (Peptides)		narrow_synonym	publication_mesh 10.1186/1471-2105-6-52 Peptides	0.25%

“WEBnma” to *data* branch



TP	FP	FN	Concept	Query	Score
Protein structure			label	description	3.37%
	Structure alignment (protein) (Protein structure alignment)		exact_synonym	publication_abstract 10.1186/s12859-014-0427-6	2.77%
	Structure		label	doc	2.51%
	~~Protein flexibility or motion report (Protein flexibility or motion)~~		exact_synonym	publication_abstract 10.1186/s12859-014-0427-6	2.50%
	Protein structure report (Protein structural property)		exact_synonym	publication_fulltext 10.1186/s12859-014-0427-6	2.38%
		Plot			
		Sequence profile			
		Structure alignment			
		Structural profile			

“WEBnma” to *format* branch



TP	FP	FN	Concept	Query	Score
	protein		label	publication_abstract 10.1186/s12859-014-0427-6	2.78%
	Format		label	doc	2.39%
	Protein secondary structure format		label	publication_abstract 10.1186/s12859-014-0427-6	1.42%
	Protein structure report (quality evaluation) format		label	publication_fulltext 10.1186/s12859-014-0427-6	1.39%
PDB			label	webpage	1.22%
		PDF			
		Textual format			
		FASTA-like (text)			

Results

Top 5 & all branches: recall 27% ; average precision 16%

Results

Top 5 & all branches: recall 27% ; average precision 16%

In other databases:	ms-utils.org:	52% ; 37%
	BioConductor:	36% ; 24%

Results

Top 5 & all branches: recall 27% ; average precision 16%

In other databases:	ms-utils.org:	52% ; 37%
	BioConductor:	36% ; 24%

Mistakes in manual annotations and tool descriptions

Results

Top 5 & all branches: recall 27% ; average precision 16%

In other databases:	ms-utils.org:	52% ; 37%
	BioConductor:	36% ; 24%

Mistakes in manual annotations and tool descriptions

And of course, deficiencies of the mapping algorithm:

Results

Top 5 & all branches: recall 27% ; average precision 16%

In other databases:	ms-utils.org:	52% ; 37%
	BioConductor:	36% ; 24%

Mistakes in manual annotations and tool descriptions

And of course, deficiencies of the mapping algorithm:

☞ inability to differentiate meaning

☞ incorrect order caused by noise & less relevant content

Results

Top 5 & all branches: recall 27% ; average precision 16%

In other databases:	ms-utils.org:	52% ; 37%
	BioConductor:	36% ; 24%

Mistakes in manual annotations and tool descriptions

And of course, deficiencies of the mapping algorithm:

☞ inability to differentiate meaning

☞ incorrect order caused by noise & less relevant content

Best metric: usefulness to the curator

Results

Top 5 & all branches: recall 27% ; average precision 16%

In other databases: ms-utils.org: 52% ; 37%
BioConductor: 36% ; 24%

Mistakes in manual annotations and tool descriptions

And of course, deficiencies of the mapping algorithm:

- ☞ inability to differentiate meaning
- ☞ incorrect order caused by noise & less relevant content

Best metric: usefulness to the curator

- ☞ many false positives make sense
- ☞ flexibility (match short and long text, tune parameters)

Conclusions

Manual annotation of thousands of bioinformatics tools

✓ very useful

✗ time-consuming and error-prone

Conclusions

Manual annotation of thousands of bioinformatics tools

- ✓ very useful
- ✗ time-consuming and error-prone

Goal to make an automatic mapper

Conclusions

Manual annotation of thousands of bioinformatics tools

✓ very useful

✗ time-consuming and error-prone

Goal to make an automatic mapper

♺ reads in free text from metadata

Conclusions

Manual annotation of thousands of bioinformatics tools

- ✓ very useful
- ✗ time-consuming and error-prone

Goal to make an automatic mapper

- ♳ reads in free text from metadata
- ♴ adds content from Internet

Conclusions

Manual annotation of thousands of bioinformatics tools

✓ very useful

✗ time-consuming and error-prone

Goal to make an automatic mapper

♳ reads in free text from metadata

♴ adds content from Internet

♵ matches against EDAM ontology terms

Conclusions

Manual annotation of thousands of bioinformatics tools

✓ very useful

✗ time-consuming and error-prone

Goal to make an automatic mapper

♳ reads in free text from metadata

♴ adds content from Internet

♵ matches against EDAM ontology terms

♶ outputs best annotation suggestions to curator

Conclusions

Manual annotation of thousands of bioinformatics tools
  ✓ very useful
  ✗ time-consuming and error-prone

Goal to make an automatic mapper
  ♳ reads in free text from metadata
  ♴ adds content from Internet
  ♵ matches against EDAM ontology terms
  ♶ outputs best annotation suggestions to curator

As result, we have a helpful curation tool

Conclusions

Manual annotation of thousands of bioinformatics tools

✓ very useful

✗ time-consuming and error-prone

Goal to make an automatic mapper

♳ reads in free text from metadata

♴ adds content from Internet

♵ matches against EDAM ontology terms

♶ outputs best annotation suggestions to curator

As result, we have a helpful curation tool

⇒ https://github.com/edamontology/edammap

Future work

Development will continue in collaboration with curators

☞ Incremental updates

  ☝ investigate effects of IDF

  ☝ customise output to curators’ needs

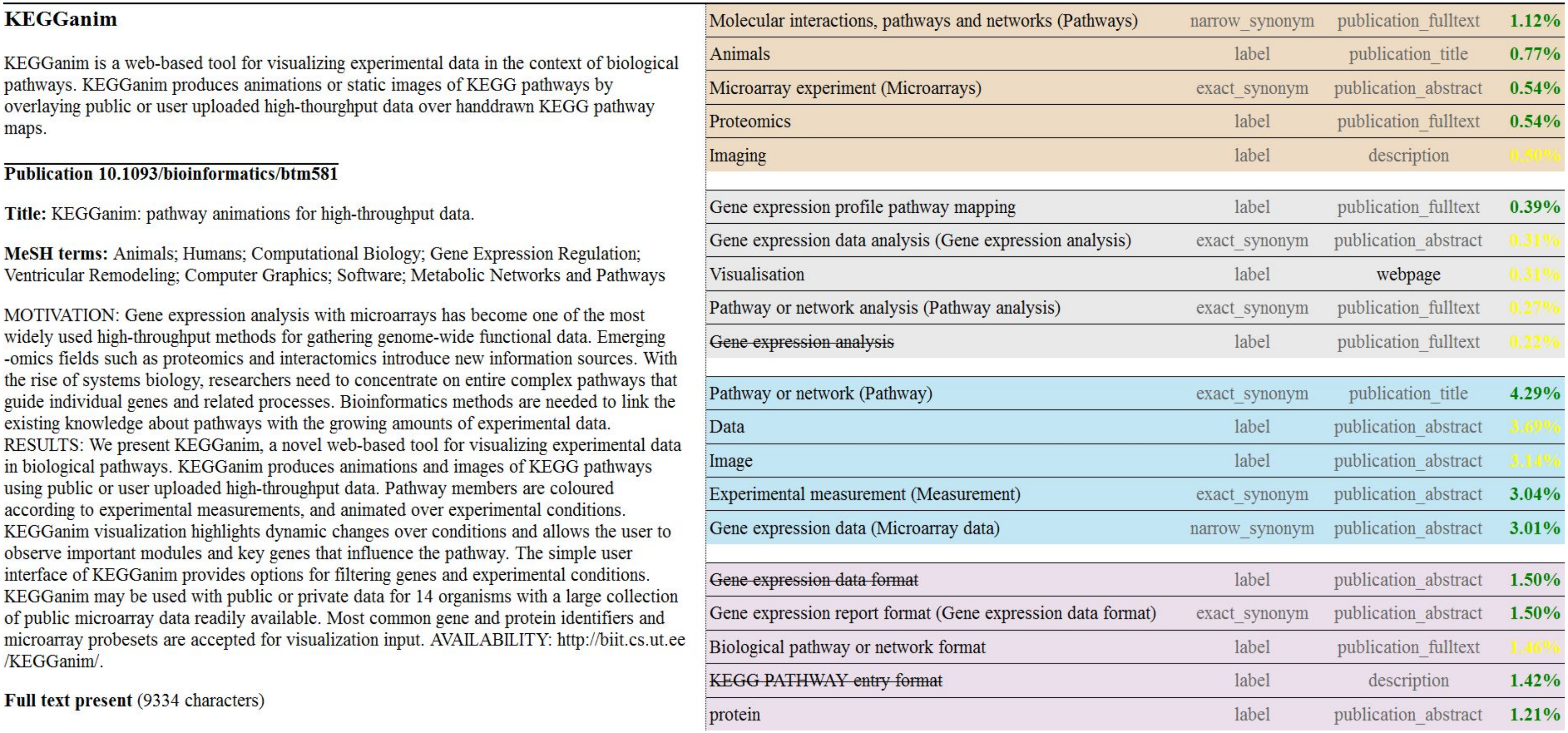
☞ Integrate with on-line bio.tools portal

☞ Annotate training materials (PPT, PDF)

☞ Discover new tools

☞ Extract new EDAM concepts

“KEGGanim” results



KEGGanim

KEGGanim is a web-based tool for visualizing experimental data in the context of biological pathways. KEGGanim produces animations or static images of KEGG pathways by overlaying public or user uploaded high-thourghput data over handdrawn KEGG pathway maps.

Publication 10.1093/bioinformatics/btm581

Title: KEGGanim: pathway animations for high-throughput data.

MeSH terms: Animals; Humans; Computational Biology; Gene Expression Regulation; Ventricular Remodeling; Computer Graphics; Software; Metabolic Networks and Pathways

MOTIVATION: Gene expression analysis with microarrays has become one of the most widely used high-throughput methods for gathering genome-wide functional data. Emerging -omics fields such as proteomics and interactomics introduce new information sources. With the rise of systems biology, researchers need to concentrate on entire complex pathways that guide individual genes and related processes. Bioinformatics methods are needed to link the existing knowledge about pathways with the growing amounts of experimental data. RESULTS: We present KEGGanim, a novel web-based tool for visualizing experimental data in biological pathways. KEGGanim produces animations and images of KEGG pathways using public or user uploaded high-throughput data. Pathway members are coloured according to experimental measurements, and animated over experimental conditions. KEGGanim visualization highlights dynamic changes over conditions and allows the user to observe important modules and key genes that influence the pathway. The simple user interface of KEGGanim provides options for filtering genes and experimental conditions. KEGGanim may be used with public or private data for 14 organisms with a large collection of public microarray data readily available. Most common gene and protein identifiers and microarray probesets are accepted for visualization input. AVAILABILITY: http://biit.cs.ut.ee /KEGGanim/.

Full text present (9334 characters)

Molecular interactions, pathways and networks (Pathways)	narrow_synonym	publication_fulltext	**1.12%**
Animals	label	publication_title	**0.77%**
Microarray experiment (Microarrays)	exact_synonym	publication_abstract	**0.54%**
Proteomics	label	publication_fulltext	**0.54%**
Imaging	label	description	**0.50%**
Gene expression profile pathway mapping	label	publication_fulltext	**0.39%**
Gene expression data analysis (Gene expression analysis)	exact_synonym	publication_abstract	**0.31%**
Visualisation	label	webpage	**0.31%**
Pathway or network analysis (Pathway analysis)	exact_synonym	publication_fulltext	**0.27%**
~~Gene expression analysis~~	label	publication_fulltext	**0.22%**
Pathway or network (Pathway)	exact_synonym	publication_title	**4.29%**
Data	label	publication_abstract	**3.69%**
Image	label	publication_abstract	**3.14%**
Experimental measurement (Measurement)	exact_synonym	publication_abstract	**3.04%**
Gene expression data (Microarray data)	narrow_synonym	publication_abstract	**3.01%**
~~Gene expression data format~~	label	publication_abstract	**1.50%**
Gene expression report format (Gene expression data format)	exact_synonym	publication_abstract	**1.50%**
Biological pathway or network format	label	publication_fulltext	**1.46%**
~~KEGG PATHWAY entry format~~	label	description	**1.42%**
protein	label	publication_abstract	**1.21%**