Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 79, No. 1 (2009) 109–129
DOI: 10.1007/s11192-009-0407-1

A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics

WOLFGANG GLÄNZEL,[a,b] FRIZO JANSSENS,[a,c] BART THIJS[a]

[a] *K.U. Leuven, Steunpunt O&O Indicatoren and Faculty ETEW, Dept. MSI, Dekenstraat 2, B-3000 Leuven, Belgium*
[b] *Institute for Research Policy Studies, Hungarian Academy of Sciences, Budapest, Hungary*
[c] *K.U. Leuven, Dept. of Electrical Engineering ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*

A novel subject-delineation strategy has been developed for the retrieval of the core literature in bioinformatics. The strategy combines textual components with bibliometric, citation-based techniques. This bibliometrics-aided search strategy is applied to the 1980–2004 annual volumes of the Web of Science. Retrieved literature has undergone a structural as well as quantitative analysis. Patterns of national publication activity, citation impact and international collaboration are analysed for the 1990s and the new millennium.

Introduction

Bioinformatics is an interdisciplinary field that emerged from the increasing use of computer science and information technology for solving problems in biomedicine, mostly at the molecular level. OUZOUNIS & VALENCIA, [2003] have provided a review of the early stages of the long history of the bioinformatics discipline. In recent studies by PATRA & MISHRA [2006] and PEREZ-IRATXETA & AL. [2006], evolution and trends in bioinformatics research have been studied. The field has been characterised as an emerging discipline with astonishing growth dynamics. The studies were based on the MEDLINE database and partially on NIH-funded project grants. In both cases, bioinformatics was analysed in a broader biomedical context. In our present paper, we will strictly focus on the *core literature* in bioinformatics. Earlier structural and dynamic analyses of the domain by JANSSENS & AL. [2006, 2007A] were strongly based on text mining and bibliometrics aided techniques, and aimed at improving classification of literature through the combination of linguistic and bibliometric tools. The aim of the present study, however, is to analyse the bibliometric core literature and its structure from the bibliometric point of view. We analyse retrieved bioinformatics

Received December 5, 2007

Address for correspondence:
WOLFGANG GLÄNZEL
E-mail: Wolfgang.Glanzel@econ.kuleuven.be

0138–9130/US $ 20.00

literature along the following research tasks. First, we will have a look at growth dynamics of the field and the sub-discipline representation of the found clusters. In a second step, national publication activity and citation impact will be studied. Finally, patterns of international co-authorship and its citation impact are analysed. Unlike the above-mentioned studies by PATRA & MISHRA [2006] and PEREZ-IRATXETA & AL. [2006], the present paper is based on literature extracted from the Web of Science of *Thomson Scientific* (part of Thomson-Reuters, Philadelphia, PA, USA). Only part of the computational linguistic analysis was conducted on MeSH terms taken from the MEDLINE database. Methodological background and data processing of this novel approach are summarised in the following sections.

Data sources and data processing

All bibliometric results are based on raw bibliographic data extracted from the 14-year annual volumes (1991–2004) of the Web of Science Edition of the Science Citation Index Expanded (SCIE) of Thomson Scientific (Philadelphia, PA, USA). Publication data have been matched with MEDLINE which has been used as auxiliary data source for the determination of search terms. Only papers recorded as *article, note* or *review* in the SCIE were taken into consideration. Papers recorded as *letter to the editor* were excluded since this document type tends to cause biases in the application of bibliographic coupling and co-citation analyses (see [GLÄNZEL & CZERWON, 1996]). The papers were assigned to countries based on the corporate address given in the by-line of the publication. All countries and institutions indicated in the address field were thus taken into account. Co-authorship was counted for the corresponding address pairs (countries and institutions) if the names of the concerning entities occurred simultaneously. It has to be stressed here that publication counts and citation frequencies cannot be summed up over co-authorship links to the total. For the meso study, addresses were cleaned-up, unified and accordingly de-duplicated at the level of main institution.

Citation counts have been determined on basis of an item-by-item procedure using special identification keys made up of bibliographic components of the author and source fields. Citations were counted in a three-year period: in the year of publication and the two subsequent years, that is, for instance, if papers published in 1999 were considered, all citations received by them in the period 1999–2001 have been counted. The choice of the citation window is in line with recent practice in the field of scientometrics. Because of the use of 3-year citation windows, citations could be counted for papers published up to 2003 (citations received in 2003–2005).

The delineation of the research field bioinformatics

An earlier bibliometric study by PATRA & MISHRA [2006] was based on MEDLINE and the use of MeSH terms resulted in a rather broad coverage of the field. In the present study we apply a much stricter strategy resulting in a *core set* of bioinformatics literature. Such *bibliometrics-aided data retrieval* strategies as described by GLÄNZEL & AL. [2006], ZITT & BASSECOULARD [2006], BASSECOULARD & ZITT [2007] have been developed for research–evaluation purposes since science policy addresses new emerging or complex interdisciplinary fields the delineation of which is particularly difficult. The objectives of subject delineation in the framework of bibliometric (domain) studies essentially differ from the goals of traditional information retrieval. In particular, bibliometrics allows including also 'metric' components in the search strategy. Thresholds of the strength of citation or bibliographic coupling links can be used to fine-tune this component. In what follows, we will present the methodological outline of bibliometric retrieval (BR) which should be understood as the extension of traditional information retrieval by bibliometric methods and applied to the delineation of subject fields. Unlike the method used by *Zitt* and *Bassecoulard* (e.g., [ZITT & BASSECOULARD, 2006]) we do not apply the bibliometric component iteratively. Normally, once the original core set is extended by bibliometric means, the resulting document set could be used as new core set for a second extension. However, the conditions used for defining the rules in fine-tuning and controlling for noise and recall tend to loose their strength through repeated application. Therefore, our retrieval "extension" is applied only once. Our search strategy actually is based on bibliographic coupling ("horizontally" searching at the same time level) as well as on references and citations ("vertically" searching in the past and future, respectively). A further data source has been used, namely the subject headings annotated to MEDLINE records that were matched with the ISI dataset. The MeSH terms are also used in part for validation and to refine the retrieval made in the SCIE database. This complex strategy applied in [JANSSENS, 2006] consists logically of two parts which, in turn, have several components each. The first part comprises two *unconditional criteria* (UC1 and UC2), which include core journals covered by the Web of Science (UC1) and the MEDLINE database (UC2), respectively.

UC1: Journal in WoS = BIOINFORMATICS (formerly COMPUTER APPLICATIONS IN THE BIOSCIENCES), JOURNAL OF COMPUTATIONAL BIOLOGY, BRIEFINGS IN BIOINFORMATICS, BMC BIOINFORMATICS.

UC2: Journal in MEDLINE = IN SILICO BIOLOGY, PSB ON-LINE PROCEEDINGS, APPLIED BIOINFORMATICS, PLOS COMPUTATIONAL BIOLOGY

UC3: Keywords in title = BIOINFORMATICS, COMPUTATIONAL BIOLOG*, SYSTEMS BIOLOGY

In other terms, all papers meeting at least one of the criteria UC1 – UC3 are considered relevant. This set has been extended by two *conditional criteria* (CC1 and CC2), each of which results in related but not necessarily in core literature. In particular, the conditional criteria comprise conditions for reference (CC2).

CC1: Records cited by UC1

CC2: Records citing UC1

All papers meeting at least one of the criteria CC1 and CC2 are considered potentially relevant, but might not directly be concerned with bioinformatics. Only that part of literature, which meets further restrictive criteria, will be considered truly relevant. In order to reduce or even exclude noise, thresholds T_i for the strength of citation and reference links were used for fine-tuning. The bibliometrics aided retrieval strategy (BR) for identifying relevant papers in bioinformatics is thus obtained by the following formula.

$$BR_{bioinf} = (UC1 \vee UC2 \vee UC3) \vee ((CC1 \wedge T_i) \vee (CC2 \wedge T_i)).$$

In particular, we used four different thresholds T_i based on the absolute number i of citations and references, respectively. Table 1 presents the effect of adjusting the strength of citation/reference links for $i = 1, 2, \ldots, 4$ on the number of retrieved documents. In addition, the results of the first unconditional criterion as well as the 'OR' combination of UC1 with the third unconditional criterion is shown. Since T_1 and T_2 still resulted in perceptible noise, we decided to use T_3 for the study.

The retrieval has first been made for the period 1981–2004; all papers indexed for the sub-period 1991–2004 have then been selected for the bibliometric analysis. All retrieval related statistics are calculated for the full 26-year time span. The publication output in the field in 1980–1990 is, however, minute and from the statistical viewpoint not decisively.

Records retrieved from WoS were matched against MEDLINE in order to obtain the Medical Subject Headings (MeSH). Matching was based on an item-by-item procedure using special identification-keys made up of bibliographic components such as publication year, volume, first page, first characters of author names and substrings of the title.

Table 1. Number of records retrieved for different combinations of criteria

Strategy	Threshold	Documents retrieved
UC1	–	3,386
UC1 ∨ UC3	–	9,620
BR	T_1	41,995
	T_2	13,239
	T_3	7,655
	T_4	5,470

Table 2. The most frequent 20 MeSH terms (excluding terms acknowledging funding)

Rank	MeSH term	Rank	MeSH term
1	Algorithms	11	Sequence Alignment/methods
2	Software	12	Proteins/chemistry
3	Humans	13	Base Sequence
4	Sequence Alignment	14	Sequence Analysis, DNA
5	Comparative Study	15	Gene Expression Profiling
6	Animals Proteins	16	Models, Genetic
7	Molecular Sequence Data	17	Internet
8	Computational Biology	18	Computer Simulation
9	Amino Acid Sequence	19	Oligonucleotide Array Sequence Analysis
10	Databases, Factual	20	Information Storage and Retrieval

Table 2 shows the 20 most frequent MeSH terms where we have excluded those terms acknowledging research support.

The TF-IDF weight (*term frequency–inverse document frequency*) was used as a relative measure to evaluate how important a term is to a document relative to the collection. In particular, the relative frequency of the term in a document is gauged against the frequency of the term in the collection. The *term frequency* is often defined as the relative frequency of a word in a document, that is, $tf = n_j/\Sigma_j n_j$. The *inverse document frequency*, in turn, is a measure of the general importance of the term. It is defined as $idf = -log (d_{T_i}/d)$, that is, the negative logarithm of the share of documents where the term T_i appears in. Finally, the TF-IDF weight is defined as the product of the two previous measures ($tfidf = tf \cdot idf$). The best 20 TF-IDF terms in titles and abstracts are presented in Table 3.

Table 3. The best 20 TF-IDF terms in titles and abstracts

Rank	TF-IDF term	Rank	TF-IDF term
1	protein	11	function
2	sequenc*	12	cluster
3	align	13	interdisciplinar*
4	gene	14	applic*
5	structur*	15	program
6	predict*	16	set
7	databas*	17	base
8	genom	18	domain
9	algorithm	19	interact*
10	model	20	famili*

Journal coverage of bioinformatics literature in the SCIE database

In total 7401 articles, notes or reviews in bioinformatics could be retrieved for the period 1981–2004. PATRA & MISHRA [2006] selected 14563 journal articles, that is, about twice as many as we have found. The main reason is the broad interpretation of bioinformatics resulting in a somewhat more liberal search strategy. The other reason is the broader coverage of the underlying database. As mentioned above, we aimed at a

very strict interpretation of the field, at retrieving the very core of bioinformatics with practically no noise. This was essential for having a solid groundwork for the cluster analysis of the retrieved literature. Nonetheless, the list of most relevant journals in bioinformatics of our exercise by and large coincides with that by Patra and Mishra. Core journals, of course, can be found at the top of the list (see Table 4).

Table 4. The 25 most frequently used journals for publishing bioinformatics literature

Rank	Journal	Frequency
1.	*Bioinformatics*	1900
2.	*Computer Applications in the Biosciences*	724
3.	*Nucleic Acids Research*	594
4.	*Journal of Computational Biology*	397
5.	*Journal of Molecular Biology*	241
6.	*Bmc Bioinformatics*	239
7.	*Genome Research*	203
8.	*PNAS USA*	189
9.	*Nature*	116
10.	*Molecular Biology and Evolution*	107
11.	*Science*	107
12.	*Protein Science*	92
13.	*Proteins-Structure Function and Genetics*	88
14.	*Protein Engineering*	84
15.	*Molecular Phylogenetics and Evolution*	63
16.	*Nature Genetics*	56
17.	*Journal of Molecular Evolution*	54
18.	*Current Opinion in Structural Biology*	51
19.	*Genomics*	46
20.	*Proteins-Structure Function and Bioinformatics*	44
21.	*Febs Letters*	37
22.	*Genome Biology*	37
23.	*Trends in Biochemical Sciences*	33
24.	*Genetics*	30
25.	*Trends in Genetics*	30

Table 5. The nine bioinformatics clusters obtained from the hybrid hierarchical cluster algorithm

Cluster	Name	Papers	Best author keyword	Best stem or phrase
1	RNA structure prediction	205	rna secondary structure	RNA
2	Protein structure prediction	1167	protein structure prediction	protein
3	Systems biology & molecular networks	694	bioinformatics	network
4	Phylogeny & evolution	749	phylogeny	phylogenet
5	Genome sequencing & assembly	640	sequencing hybridization	base sequenc
6	Gene/promoter/motif prediction	995	gene regulation	gene
7	Molecular DBs & annotation platforms	1091	genome analysis	databas
8	Multiple Sequence alignment	713	sequence alignment	align
9	Microarray analysis	1147	microarray	microarrai
Total	Bioinformatics	7401	bioinformatics	protein

Journals in computational and molecular biology as well as the important multidisciplinary journals *PNAS USA*, *Nature* and *Science* are the most important publication channels for bioinformatics research. The huge number of journals in which the papers were scattered in the paper by Patra and Mishra could be confirmed by us as well.

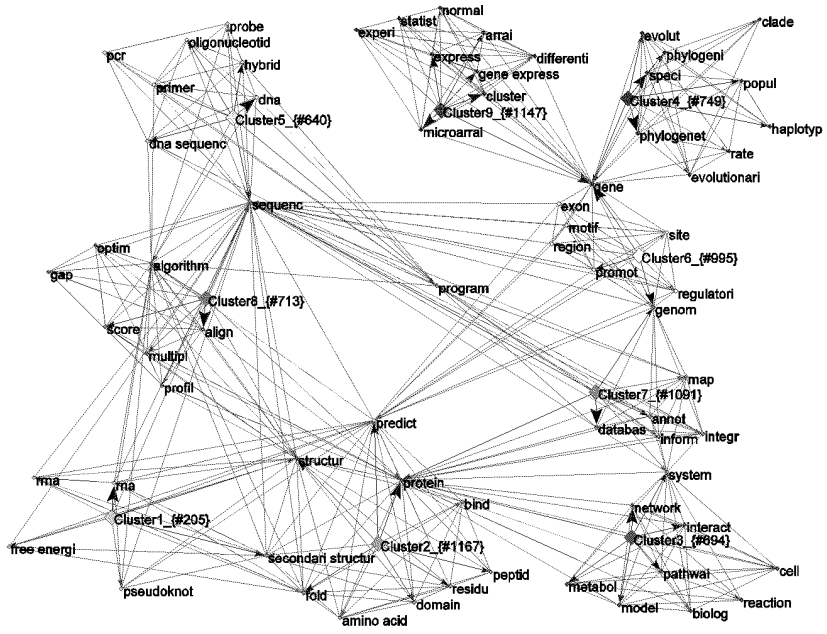

Figure 1. Visualisation of the cognitive structure of bioinformatics using Pajek [BATAGELJ, 2002]. The structure is based on term networks

In order to depict the cognitive structure of the field represented by its core literature, the agglomerative hierarchical, hard cluster algorithm using Ward's method was used (cf., [JAIN & DUBES, 1988; BERKHIN, 2002; KAUFMAN & ROUSSEEUW, 1990]). The hybrid algorithm, which is based on the integration of both textual information and citation links, is described by JANSSENS & AL. [2005, 2007A, 2007B]. In total, we obtained nine clusters. The cognitive structure of the field as reflected by term networks using the best 10 terms from titles and abstracts according to mean TF-IDF scores for each of nine clusters is shown in Figure 1. In addition, Table 5 presents the clusters, their size and their characterisation by best author keywords and best terms from titles and abstracts.

Evolution of publication output and citation impact in bioinformatics in the period 1991–2004

Evolution of publication output and citation impact of the field

Figure 2 visualises the evolution of the cumulative number of papers in bioinformatics. The growth of publications lies in between the linear model in the first half of the period and the exponential model for the second half (similarly as observed in nanoscience and -technology, [GLÄNZEL & AL., 2003]). Literature growth clearly characterises the field as a young, emerging and dynamically evolving discipline.

The dynamic growth of literature in bioinformatics is outrun by an even more powerful increase of citations. The patterns are shown in Figure 3. Citations, as already mentioned in the outset, have been determined on a basis of three-year citation windows. Before we have a closer look at citation patterns, we will introduce the indicators used for the analysis.

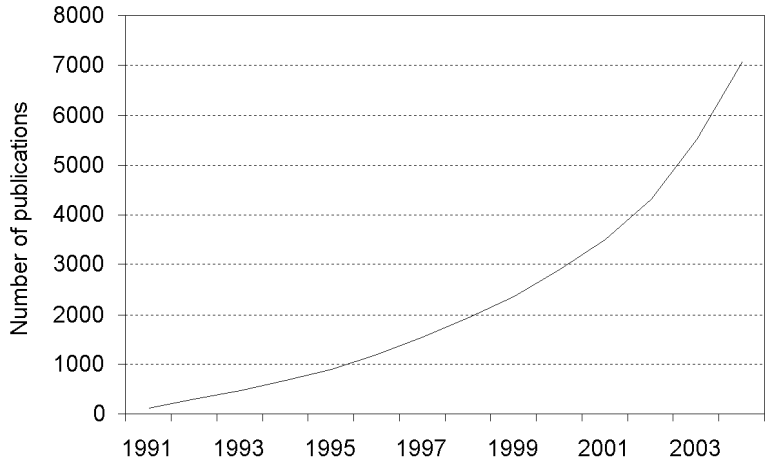

Figure 2. Evolution of cumulated publication output in the period 1991–2004 (world total)

The *Mean Observed Citation Rate* (MOCR) is defined as the ratio of citation count to publication count. It reflects the factual citation impact of any unit like a country, region, institution, research group etc. Since the underlying paper set is restricted to a single, however cross-disciplinary subject, we can use the subject-standardised Mean Observed Citation Rate ($MOCR|_f$) which is simply the ratio of the unit's MOCR value

and the world standard of the field. In addition, we use the share of author self-citations and the citation impact of internationally co-authored papers.

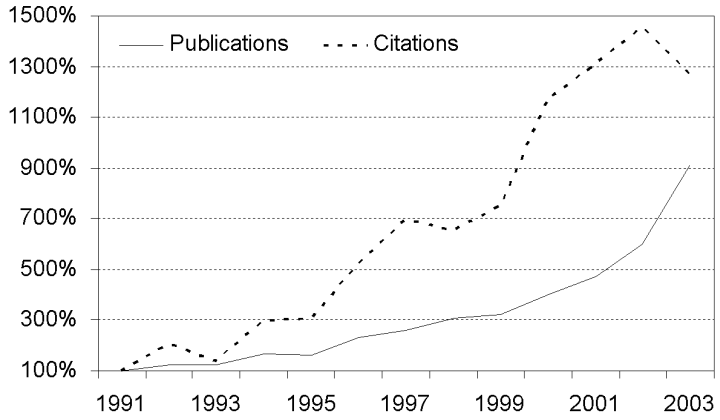

Figure 3. Annual change of citations compared with that of publications in bioinformatics for 1991–2003 (1991 = 100%)

The evolution of the field's mean observed citation impact is presented in Figure 4. The strong linear increase of citation impact in the 1990s is followed by a sharp decline in the new millennium. The reasons for this phenomenon are not clear. However, a decline of citation impact has also been observed in the case of nanoscience and -technology [GLÄNZEL & AL., 2003]. It seems that emerging fields are characterised first by a growth of citation impact exceeding that of the publication output, then by stagnation and later on by a decrease of impact while the powerful increase of publication activity continues.

In order to gain detailed information about the evolution of citation means, we analyse the distributions of citations over individual papers, one each for the beginning and the end of the period under study. The diagram is presented in Figure 5.

Although the citation impact decreases from 2001, the MOCR values for the second sub-period are still distinctly higher than the corresponding values for the first one. The distributions for 1991–1995 and 2000–2003 are quite similar except for the shares of poorly and frequently cited papers. The distribution has evolved into a slightly less skewed one. The moments of the two distributions are high: The mean in 1991–1995 amounts to 22.2, that in 2000–2003 to 31.1. The share of less cited and uncited papers decreased while that of frequently cited papers increased. In verbal terms, the frequency

distributions of citations over publications characterise the field as a specialty with high citation impact; however, the citation patterns are quite polarised although in the second sub-period less distinctly.

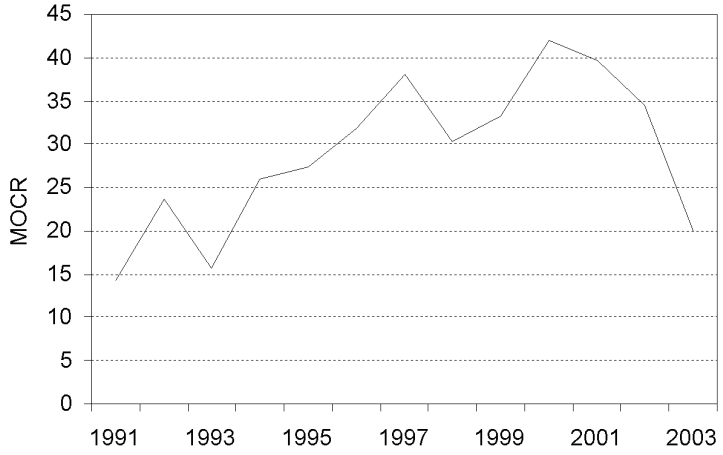

Figure 4. Evolution of mean observed citation impact in the period 1991–2003 (world total)

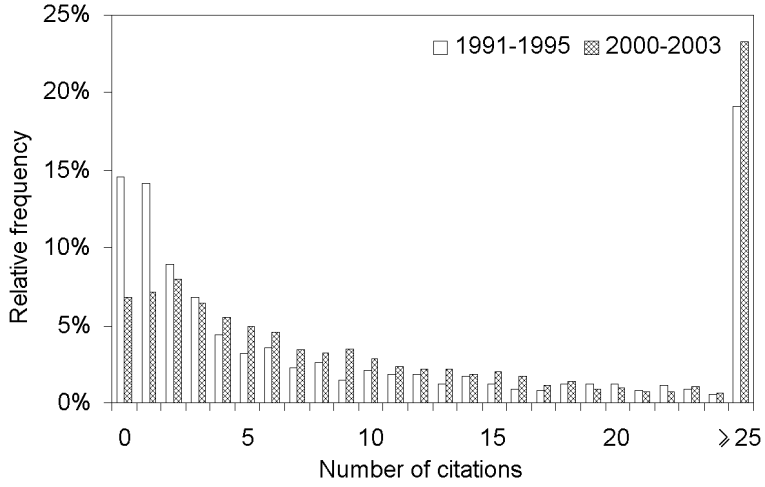

Figure 5. Distribution of citations over papers published in 1991–1995 and 2000–2003

Table 6 presents the cluster size and citation impact of the nine clusters for the period 1991–2003. Citations have been collected for 3-year time windows each

(beginning with the publication year). Cluster #1 labelled "RNA structure prediction" with 152 publications is the smallest one; all other clusters have more than 400 and less than 1000 papers. The citation impact of most clusters lies around that of the field of about 30 citations per paper. The impact of Clusters #5 (Genome sequencing & assembly) and #8 (Multiple Sequence alignment) lies far below, that of Cluster #6 (Gene/promoter/motif prediction) distinctly above the field standard.

Table 6. Publication output and citation impact of the nine bioinformatics clusters for 1991–2003

Cluster	Name	Papers	MOCR
1	RNA structure prediction	152	24.22
2	Protein structure prediction	923	27.05
3	Systems biology & molecular networks	476	26.89
4	Phylogeny & evolution	480	29.92
5	Genome sequencing & assembly	432	11.78
6	Gene/promoter/motif prediction	781	47.24
7	Molecular DBs & annotation platforms	907	33.18
8	Multiple Sequence alignment	558	16.60
9	Microarray analysis	798	37.06
Total	Bioinformatics	5507	30.27

Publication output and citation impact of the 30 most active countries

For the analysis of national publication activity and citation impact, the 30 most active countries in the period 1991–2004 have been selected. Analogously to earlier studies (e.g. [REIST-2, 1997]), a so-called full–counting or integer–counting scheme was applied, that is, a full count was recorded whenever a country occurred in the corporate address field. Duplicates have been removed. Because of the extensive presence of international co-authorship, national bibliometric indicators such as publication or citation counts based on this full-counting scheme are not additive, that is, they can not be summed up over countries to regions or supra-national units. Consequently, a share of x% of a given country in the world's total publication output means that x% all papers have one or more co-authors with an address in this country.

Countries with less than 30 papers in the 14-year period have not been selected by reasons of statistical reliability. The publication output of the 30 most active countries in bioinformatics and their share in the world total in this field are presented in Table 7. In order to provide information about the evolution of national publication activity in the field, the period 1991–2004 has been split into two sub-periods, particularly, 1991–1997 and 1998–2004. National data in Table 7 are ranked in descending order by their publication output in the whole 14-year period. If we compare the list with similar lists on national publication output in all fields combined, we can conclude that those countries that are most active in scientific research in all fields combined have top activity in bioinformatics research, too.

Table 7. Publication output of the 30 most active countries in sub-periods 1991–1997 and 1998–2004

Country	1991–1997		1998–2004		1991–2004	
	Papers	Share	Papers	Share	Papers	Share
USA	721	46.8%	2923	52.8%	3644	51.5%
GBR	235	15.3%	767	13.9%	1002	14.2%
DEU	189	12.3%	594	10.7%	783	11.1%
FRA	121	7.9%	331	6.0%	452	6.4%
JPN	74	4.8%	232	4.2%	306	4.3%
CAN	49	3.2%	223	4.0%	272	3.8%
ITA	60	3.9%	150	2.7%	210	3.0%
ESP	39	2.5%	146	2.6%	185	2.6%
ISR	33	2.1%	144	2.6%	177	2.5%
SWE	14	0.9%	161	2.9%	175	2.5%
RUS	56	3.6%	118	2.1%	174	2.5%
AUS	21	1.4%	134	2.4%	155	2.2%
CHE	47	3.1%	100	1.8%	147	2.1%
CHN	7	0.5%	139	2.5%	146	2.1%
BEL	24	1.6%	108	2.0%	132	1.9%
DNK	12	0.8%	83	1.5%	95	1.3%
NLD	18	1.2%	77	1.4%	95	1.3%
IND	16	1.0%	72	1.3%	88	1.2%
SGP	6	0.4%	73	1.3%	79	1.1%
POL	5	0.3%	53	1.0%	58	0.8%
NOR	6	0.4%	45	0.8%	51	0.7%
IRE	7	0.5%	43	0.8%	50	0.7%
TWN	1	0.1%	47	0.8%	48	0.7%
AUT	5	0.3%	42	0.8%	47	0.7%
FIN	5	0.3%	41	0.7%	46	0.7%
KOR	1	0.1%	44	0.8%	45	0.6%
BRA	0	0.0%	44	0.8%	44	0.6%
NZL	6	0.4%	36	0.7%	42	0.6%
HUN	11	0.7%	27	0.5%	38	0.5%
GRC	5	0.3%	30	0.5%	35	0.5%
WORLD	1540	100.0%	5536	100.0%	7076	100.0%

However, the three "leading" countries, USA, UK and Germany rank distinctly higher in bioinformatics than in all fields combined (cf., [GLÄNZEL & AL., 2002]). Although publication counts of most countries for the first period are small, we can observe the same powerful growth of publication activity of China and other emerging scientific nations like South Korea, Taiwan and Brazil (see [GLÄNZEL & AL., 2008]). National representation also confirms the findings by PATRA & MISHRA [2006].

The citation impact of the 30 most active countries with at least 25 papers in 1991–2003 in the two sub-periods 1991–1997 and 1998–2003 is shown in Table 8. The overall high impact is partially a consequence of the citation-based component of the retrieval strategy. A study of bibliographic coupling by GLÄNZEL & CZERWON [1996] has shown that retrieval based on strong coupling links results on higher-than-average citation impact. Citation aided tools in information retrieval and data mining necessarily imply a certain bias concerning visibility of the literature. The better depiction of the structure of the information space is to the detriment of loosely linked and less visible documents.

Table 8. Citation impact and self-citation rate of the 30 most active countries in 1991–2003 in the two sub-periods 1991–1997 and 1998–2003

Country	1991–1997			1998–2003			1991–2003		
	Papers	MOCR|$_f$	f_S	Papers	MOCR|$_f$	f_S	Papers	MOCR|$_f$	f_S
USA	721	1.28	10.1%	2162	1.37	9.1%	2883	1.35	9.3%
GBR	235	1.17	12.0%	594	1.47	11.0%	829	1.39	11.2%
DEU	189	1.24	13.8%	429	1.48	11.2%	618	1.41	11.8%
FRA	121	2.09	12.5%	247	1.66	9.6%	368	1.78	10.6%
JPN	74	1.01	17.3%	157	1.97	10.9%	231	1.68	12.0%
CAN	49	2.96	11.1%	140	2.15	10.1%	189	2.34	10.4%
ITA	60	0.90	19.4%	103	0.73	19.9%	163	0.78	19.7%
RUS	56	0.16	26.7%	94	0.52	17.6%	150	0.39	18.9%
ISR	33	0.40	21.5%	112	2.06	9.1%	145	1.73	9.7%
ESP	39	1.20	17.5%	99	2.18	10.3%	138	1.93	11.5%
SWE	14	–	–	105	1.63	9.5%	119	1.85	8.8%
CHE	47	2.24	12.0%	68	3.04	8.6%	115	2.69	9.7%
AUS	21	–	–	90	2.49	8.9%	111	2.13	9.2%
BEL	24	–	–	71	0.88	14.2%	95	1.14	17.5%
CHN	7	–	–	79	1.96	8.9%	86	1.90	9.2%
DNK	12	–	–	61	1.64	8.0%	73	1.78	8.5%
NLD	18	–	–	47	2.56	8.5%	65	2.42	10.5%
IND	16	–	–	42	0.29	19.4%	58	0.23	20.1%
SGP	6	–	–	42	0.59	22.9%	48	0.54	23.2%
NOR	6	–	–	35	1.65	10.4%	41	1.50	10.9%
POL	5	–	–	34	0.52	27.3%	39	0.69	25.2%
IRE	7	–	–	30	4.31	7.2%	37	4.40	9.1%
FIN	5	–	–	31	0.56	13.3%	36	0.55	13.9%
HUN	11	–	–	22	–	–	33	0.42	16.8%
NZL	6	–	–	24	–	–	30	0.83	13.3%
AUT	5	–	–	24	–	–	29	0.60	17.6%
BRA	0	–	–	27	0.26	32.9%	27	0.26	32.9%
GRC	5	–	–	21	–	–	26	2.33	10.6%
TWN	1	–	–	25	0.21	26.7%	26	0.21	28.0%
KOR	1	–	–	20	–	–	21	–	–
WORLD	1540	1.00	11.3%	3967	1.00	10.2%	5507	1.00	10.5%

The high relative citation impact of Canada, Switzerland, Australia and the Netherlands (more than twice the world standard) is worth mentioning. This is contrasted by the relatively low impact of Russia and Italy in all sub-periods although their publication activity is quite high. The share of author self-citations f_S of about 10% is low in this field; national deviation from this standard follows the patterns observed from other science fields [GLÄNZEL & AL., 2003]. In general, self-citation rates are rather low, even for a biomedicine related field.

Cluster representation of the six most active countries

The breakdown of national publication output by clusters does not allow any reliable quantitative analysis for most of the 30 selected countries because of the often too small publication sets. We restrict the analysis to the six leading countries, particularly, the USA, UK, Germany, France, Japan and Canada. Their share in the nine individual

clusters is presented in Figure 6. The US with share of about 50% and more are predominant in most sub-disciplines. Above all, Cluster #9 (Microarray analysis) is dominated by the USA with 70% of all papers. Also Germany has a well-balanced high share in all clusters, except for Cluster #9. The other three countries reflect a rather heterogeneous picture; the UK contribution to Cluster #2 (Protein structure prediction) and #6 (Gene/promoter/motif prediction) is worth mentioning, however, the contribution to Cluster #1 (RNA structure prediction) and #9 (Microarray analysis) is rather small. The situation in France is similar: the strong contribution to Cluster #1 and #7 (Molecular DBs & annotation platforms) is contrasted by a low share in Cluster #9. The extremes in the Japanese publication output can be found in Cluster #3 (Systems biology & molecular networks) with 7% of the world total and Cluster #5 (Genome sequencing & assembly) with 1%. The results of further analysis of cluster dynamics and structural changes will be the subject of a forthcoming study. The situation in Canada differs from profiles of the previous countries; here the extremes can be found in Cluster #2 (Protein structure prediction) and #4 (Phylogeny & evolution).

The above presentation can advantageously be supplemented by a cluster-wise cross-national comparison which should be independent of the national publication output but which requires, of course, a certain national minimum activity as well. For this purpose we use the *Activity Index* (AI). This index, which, in turn, is a version of the economists' Comparative Advantage Index, was originally introduced by FRAME [1977]. AI is long used as relative indicator in bibliometrics (see [SCHUBERT & BRAUN, 1987]).

In our context we can, for instance, define AI as

$$AI = \frac{\text{the share of the given cluster in all publications of the given country in the field}}{\text{the share of the given cluster in the world's total publication output in the field}}$$

In verbal terms, the indicator expresses the relative effort in the individual topics (subfields) as compared with the world average.

AI's neutral value is 1.0, that is, the national activity in a subfield defined by a given cluster is in line with the world standard if AI = 1. AI = 0 indicates a completely idle subfield, AI < 1 indicates a lower-than-average, AI > 1 a higher-than-average activity. It should be noted that AI reflects a certain internal balance among the subfields in the given country, i.e., AI values greater then 1.0 must always be balanced by values less than 1. In no country can all AI values be above (or below) the world standard.

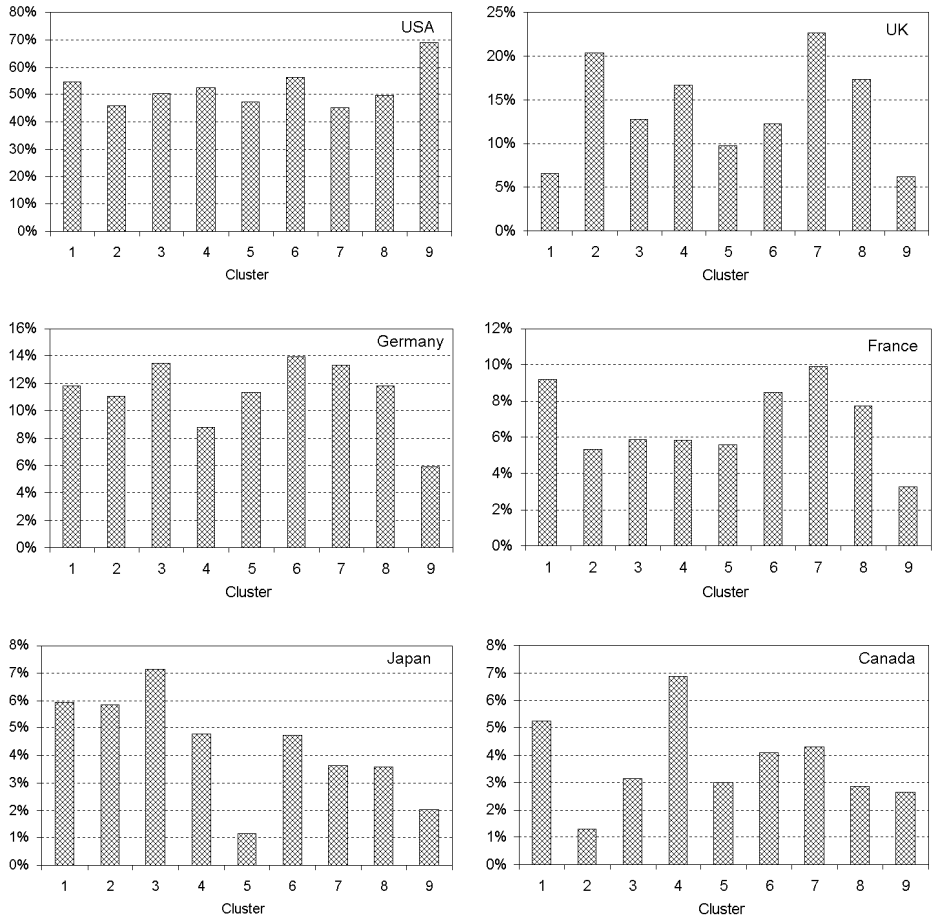

Figure 6. National representation of the six most active countries by clusters

The values of national relative publication activity are preferably presented in 'spider' diagrams although, because of the disjoint clusters, national cluster presentation forms a real distribution in our case. The diagrams for the six most active countries are presented in Figure 7.

The US profile in bioinformatics research is very close to the world standard indicated by the dotted line in the diagram. This is quite natural if one recalls that about 50% of the world's publication output in this field have an author from the USA.

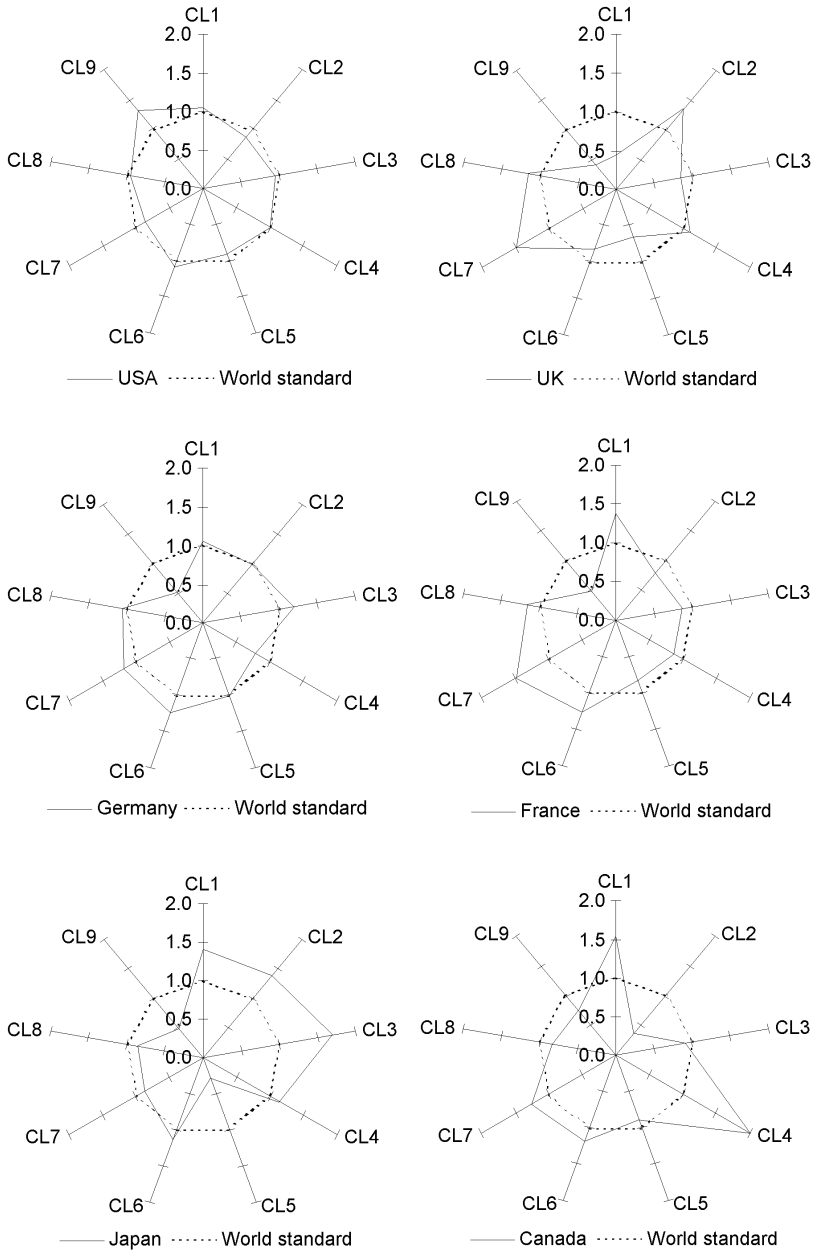

Figure 7. National subject profiles of the six most active countries according to clusters based on the Activity Index

Nevertheless, Cluster #9, where the USA has the lion's share of the world total, lies distinctly above the world standard. The six selected countries have completely different activity patterns: Germany is close to that of the USA, but with pronouncedly lower relative activity in Cluster #9. The diagrams of the other selected countries, above all those of the UK and Japan with very high relative activity in Cluster #2, #7 and Clusters #1-3, respectively, reflect a more polarised activity.

International co-authorship patterns in bioinformatics

The global collaboration network of research in bioinformatics

Beyond individual interests and motivation of individual scientists, teamwork and scientific collaboration is one of the characteristics of "big science" [PRICE, 1966]. Of course, in inter- and cross-disciplinary areas, where scientists from different fields are jointly doing research, intensive collaboration is expected (see [GLÄNZEL & AL., 2003]).

It is clear that a variety of different purposes and motivations, the manifold of factors influencing (international) collaboration, must have at least in part a measurable impact on the published results of joint research work. National characteristics in international scientific co-authorship patterns have been studied by GLÄNZEL [2001]. The results often confirmed but sometimes contradicted widespread notions on the efficiency of international collaboration. Furthermore, an interesting observation has been made concerning the re-integration of EIT countries into the scientific collaboration structures of Europe and the Western world.

The absolute number of international papers and their share in the total national publication output serve as basic indicators of international co-authorship relations and scientific collaboration. International collaboration depends on the country's 'size' (cf., for instance, [SCHUBERT & BRAUN, 1990] and [KATZ, 2000]). At the national level, the share of international collaboration in large countries is necessarily lower than that of medium-sized or even small countries. The share of all international papers in the world can, in principle, be determined as the complementary share of the ratio of all countries' domestic papers and the total world publication output. Such 'world average' is, however, not an appropriate reference standard for international collaboration activity (cf. [SCHUBERT & BRAUN, 1990]), and is therefore not used here.

Table 9 presents number and share of internationally co-authored publications of those of the 19 most active countries that have at least 25 international papers each in the period 1991–2004. Countries have been ranked by the share of international co-publications in the total national publication output. In addition, both the national MOCR values and the corresponding indicator for international co-publications ($MOCR|_i$) are presented.

Hungary, the Netherlands and Denmark have the highest share of international co-publications. More than two thirds of their papers have been published in international collaboration. Among the countries with high share of international co-publications, we also find Israel, Sweden, Singapore, China, Switzerland, Canada and Russia with more than 50% international papers. Even US scientists publish one quarter of their papers jointly with colleagues abroad. In all, the shares of international co-publications are roughly in line with those found in other interdisciplinary fields like, for instance, nanoscience and -technology (cf., [GLÄNZEL & AL., 2003]).

Table 9. Share and citation impact of international co-publications of 19 selected countries in 1991–2003

Country	Co-publ.	Share	MOCR	$MOCR\|_i$
HUN	25	75.8%	12.64	11.92
NLD	46	70.8%	73.38	97.02
DNK	49	67.1%	53.90	73.20
ISR	85	58.6%	52.41	80.12
SGP	28	58.3%	16.38	24.11
SWE	68	57.1%	55.93	89.96
CHN	49	57.0%	57.38	95.55
CHE	65	56.5%	81.57	113.18
CAN	97	51.3%	70.91	116.03
RUS	76	50.7%	11.78	20.42
AUS	55	49.5%	64.44	120.73
ESP	68	49.3%	58.41	106.53
BEL	43	45.3%	34.37	56.35
ITA	70	42.9%	23.63	49.37
DEU	265	42.9%	42.63	73.43
GBR	324	39.1%	42.05	78.02
FRA	139	37.8%	53.77	120.79
JPN	73	31.6%	50.98	131.66
USA	707	24.5%	40.91	61.53

The citation indicators are even more impressing. The figures confirm that international collaboration in general results in higher visibility and impact, but as mentioned above, there are also exceptions to the rule. The already very high citation scores are outrun by the reception of the international papers in our set. Almost incredible values are reached by Switzerland, Canada, Australia, Spain, France and Japan. On the other hand, collaboration seems not to pay off for Hungary; despite the huge share of collaboration, domestic papers have a better reception here.

Mapping mutual co-authorship links

In order to measure the strength of mutual collaboration links, an appropriate similarity measure based on country pairs is used. Multinational collaboration is therefore split up to a group of bilateral relations. In particular, binary links between the countries are studied. A link between two countries is established whenever the two

given countries under study co-occurred in the corporate address in the by-line of a publication. In this context it should be mentioned that as a consequence of treating collaboration links of each country pair separately, co-publication counts and shares are not additive, and therefore cannot be summed up to the total over any part of the world. One has, consequently, to distinguish between the number of co-publications and of co-authorship links.

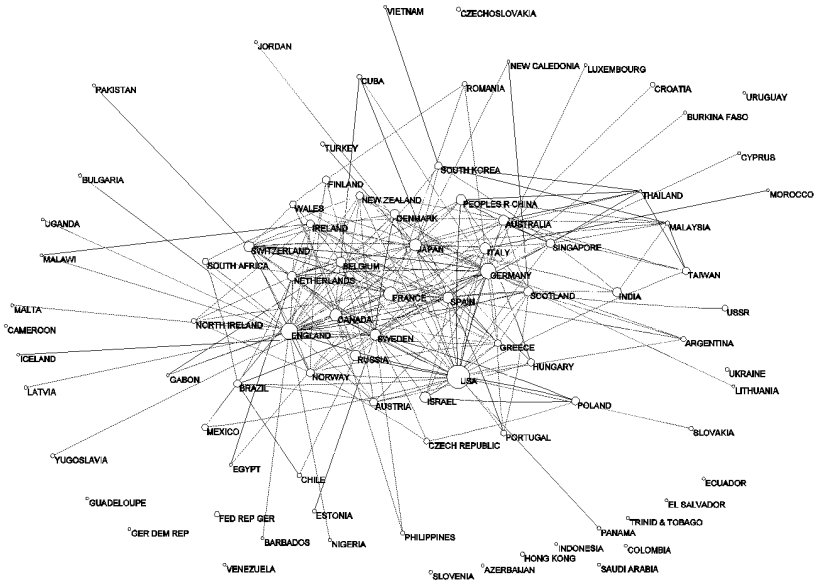

Figure 8. International collaboration network based on Salton's cosine measure with Kamada-Kawai layout [BATAGELJ, 2002]

A frequently used measure for the strength of co-publication links is the cosine measure according to Salton. It is defined as the number of joint publications divided by the square root of the product of the number (i.e., the geometric mean) of total publication outputs of the two countries, that is,

$$r = \frac{p_{ij}}{\sqrt{p_i \cdot p_j}},$$

where p_{ij} is the number of links between the countries i and j, and p_i (p_j) the total number of publications of the country i (j). As a consequence of this practice one has to distinguish between the number of *co-publications* and of *co-authorship links*.

The results are presented in Figure 8. The "big" countries, USA, UK, Germany, France and Japan can be found in the very centre of the diagram. These countries are the real nodes of this global network. Since the figure is based on all bioinformatics papers retrieved for 1980–2004, countries like Czechoslovakia, GDR and FRG still appear in the diagram; because of the dynamical growth of the field, their role in the complete set is marginal. The appearance of the emerging nations like China, Singapore, Korea and Brazil as nodes in the collaboration network is again worth mentioning.

Conclusions

The field of bioinformatics proved a young, emerging field characterised by a powerful, from the late 1990s on, by an almost exponential growth of literature. Beyond several core journals, important periodicals in molecular biology as well as the multidisciplinary journals *Science*, *Nature* and *PNAS USA* proved to be the most important publication channels. Although we focussed on the bioinformatics core literature, our study has confirmed findings by other recent studies concerning publication patterns.

The structural analysis resulted in the identification of nine sub-disciplines with individual national profiles. The partially citation-based subject delineation supported the identification of rather visible publications; the citation analysis characterised bioinformatics as a field with very high overall citation scores. According to our expectations, the extent of international collaboration is in keeping with that of other emerging interdisciplinary fields. The "big" countries form the nodes of the global co-publication network. International collaboration resulted in general to a powerful increase of the otherwise already high citation impact.

*

An extended version of a paper presented at the *11th International Conference on Scientometrics and Informetrics*, Madrid (Spain), 25-27 June 2007 [GLÄNZEL & AL., 2007]

References

BASSECOULARD, E., LELU, A., ZITT, M. (2007), A modular sequence of retrieval procedures to delineate a scientific field: from vocabulary to citations and back. In: E. TORRES-SALINAS, H. F. MOED (Eds), *11th International Conference on Scientometrics and Informetrics (ISSI 2007)*, Madrid, Spain, 25-27 June 2007, 74–84.

BATAGELJ, V., MRVAR, A. (2002), Pajek – Analysis and visualization of large networks, *Graph Drawing*, 2265 : 477–478.

BERKHIN, P. (2002), *Survey of Clustering Data Mining Techniques*, Technical report (Accrue Software). Retrieved November 15, 2006 from: http://citeseer.ist.psu.edu/berkhin02survey.html.

FRAME, J. D. (1977), Mainstream research in Latin America and the Caribbean, *Interciencia*, 2 : 143–148.

GLÄNZEL, W., CZERWON, H. J. (1996), A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level, *Scientometrics*, 37 : 195–221.

GLÄNZEL, W. (2001), National Characteristics in International Scientific Co-authorship, *Scientometrics*, 51 : 69–115.

GLÄNZEL, W., SCHUBERT, A., BRAUN, T. (2002), A relational charting approach to the world of basic research in twelve science fields at the end of the second millennium, *Scientometrics*, 55 (3) : 335–348.

GLÄNZEL, W., MEYER, M., DU PLESSIS, M., THIJS, B., MAGERMAN, T., SCHLEMMER, B., DEBACKERE, K., VEUGELERS, R. (2003), *Nanotechnology – Analysis of an Emerging Domain of Scientific and Technological Endeavour*, Retrieved November 15, 2006 from: http://www.steunpuntoos.be/nanotech_domain_study.pdf

GLÄNZEL, W., THIJS, B., SCHLEMMER, B. (2004), A bibliometric approach to the role of author self-citations in scientific communication, *Scientometrics*, 59 : 63–77.

GLÄNZEL, W., JANSSENS, F., SPEYBROEK, S., SCHUBERT, A., THIJS, B. (2006), *Towards a Bibliometrics-Aided Data Retrieval for Scientometric Purposes,* Poster presented at the 9th International Conference on Science and Technology Indicators, Leuven (Belgium), 7-9 September 2006.

GLÄNZEL, W., JANSSENS, F., THIJS, B. (2007), A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics, In: D. TORRES-SALINAS, H. MOED: *Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics*, 221–234.

GLÄNZEL, W., DEBACKERE, K., MEYER, M. (2008), 'Triad' or 'Tetrad'? On global changes in a dynamic world, *Scientometrics*, 74 : 71–88.

JANSSENS, F., GLENISSON, P., GLÄNZEL, W., DE MOOR, B. (2005), Co-clustering approaches to integrate lexical and bibliographical information, In: P. INGWERSEN, B. LARSEN (Eds), *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, pp. 284–289. Stockholm, Sweden.

JANSSENS, F., TRAN QUOC, V., GLÄNZEL, W., DE MOOR, B. (2006), Integration of textual content and link information for accurate clustering of science fields, In: V. P. GUERRERO-BOTE (Ed.), *Proceedings of the I International Conference on Multidisciplinary Information Sciences and Technologies (InSciT2006)*, pp. 615–619. Mérida, Spain.

JANSSENS F. (2007A), *Clustering of Scientific Fields by Integrating Text Mining and Bibliometrics*, Ph.D. thesis, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium, http://hdl.handle.net/1979/847.

JANSSENS, F. GLÄNZEL, W., DE MOOR, B. (2007B), Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis, In: P. BERKHIN, R. CARUANA, X. WU, S. GAFFNEY (Eds), *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*, pp. 360–369. San Jose, California, USA. ACM Press.

JAIN, A., DUBES, R. (1988), *Algorithms for Clustering Data*, New Jersey: Prentice Hall.

KATZ, J. S. (2000), Scale-independent indicators and research evaluation, *Science and Public Policy*, 27 (1) : 23–36.

KAUFMAN, L., ROUSSEEUW, P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: John Wiley and Sons Inc.

OUZOUNIS, C. A., VALENCIA, A. (2003), Early bioinformatics: the birth of a discipline – a personal view, *Bioinformatics*, 19 : 2176–2190.

PATRA, S. K., MISHRAM S. (2006), Bibliometric study of bioinformatics literature, *Scientometrics*, 67 : 477–489.

PEREZ-IRATXETA, C., ANDRADE-NAVARRO, M. A., WREN, J. D. (2006), Evolving research trends in Bioinformatics, *Briefings in Bioinformatics*, in press.

PRICE, D. DESOLLA (1966), *Little Science, Big Science*, New York: Columbia Univ. Press.

REIST-2 (1997), *The European Report on Science and Technology Indicators 1997*, EUR 17639. European Commission. Brussels.

SCHUBERT, A. BRAUN, T. (1986), Relative indicators and relational charts for comparative assessment of publication output and citation impact, *Scientometrics*, 9 (5–6) : 281–291.

SCHUBERT, A., BRAUN, T. (1990), World flash on basic research: International collaboration in the sciences, 1981–1985, *Scientometrics*, 19 : 3–10.

ZITT, M., BASSECOULARD, E. (2006), Delineating complex scientific fields by hybrid lexical-citation method: an application to nanoscience, *Information Processing & Management*, 42 (6) : 1513–1531.