Good Practice in Traditional Chinese Medicine Research in the Post-Genomic Era

GP-TCM

223154

D4.12

Proceedings of the interdisciplinary consensus meeting on *in-silico* tools for CHM research

Document description	
Name of document	Proceedings of the interdisciplinary consensus meeting on *in-silico* tools for CHM research
Abstract	Discussions held between an interdisciplinary team comprising WP4 members and a specially co-opted *in-silico* work-group were used as the basis for deriving criteria for evaluation of CHM databases and *in-silico* analytical tools. The criteria for CHM databases are used in evaluating the currently available databases. Some general observations are presented to guide use of *in-silico* tools in CHM research.
Document identifier	D4.12
Document class	Deliverable
Version	1
Author(s)	David Barlow, Tom Ehrman, Alessandro Buriani
Date of creation	21st April 2011
Date of last modification	30th April 2011
Status	Final
Destination	European Commission
WP number	WP4

TABLE OF CONTENTS

1 INTERDISCIPLINARY CONSENSUS MEETING ON *IN-SILICO* TOOLS FOR CHM RESEARCH

1.1 Introduction

In order to expedite a timely development of consensus in regard to the utility of *in-silico* tools and databases in CHM research, whilst at the same time ensuring that all interested and informed parties were able to make a full contribution, it was deemed appropriate to reduce the scale of the meeting – and thus to involve only key WP4 members – but with the meeting preceded by an extensive consultation by e-mail. In anticipation of this need, the questionnaire on "Functional genomics techniques for *in-vitro* CHM research", prepared under WP4 D4.14, was extended to include an additional two questions dealing specifically with *in-silico* CHM research. This questionnaire (presented in section 1.3, D4.14) was duly circulated to all WP4 members, and the answers collected and analysed at the WP4 D4.12 consensus meeting, held in London, 31st March – 1st April 2011.

Note too that, given an early realisation that the WP4 membership included only a rather limited number of personnel with expertise in *in-silico* tools and databases, it was also deemed necessary to co-opt a number of additional members who would be able to contribute specifically under the WP4 deliverables D4.12 - D4.14. These co-opted members – convened as the WP4 *in-silico* work-group – included Dr Thomas Ehrman (King's College London), Dr Ivano Eberini (University of Milan), Dr Judith Rollinger (University of Innsbruck), and Professor Weiliang Zhu (Shanghai Institute of Materia Medica).

The views and opinions of the *in-silico* work-group were sought *via* e-mail - as noted in the various sections below - and the work-group members were also circulated with the "Functional genomics …." questionnaire referred to above.

The combined WP4 membership and *in-silico* work-group thus contributed to the drafting of the consensus meeting's conclusions *via* e-mail and/or questionnaire responses and/or direct participation.

1.2 Discussion on terminology and other communication problems between specialist and non specialist

Preliminary discussions at the London *in-silico* meeting focussed on the problems of terminology, and the difficulties associated with performing scientifically sound and fruitful *in-silico* studies in CHM research. For the purposes of the meeting, *in-silico* studies were defined as those involving virtual screening and/or cheminformatics. The view was universally held that

such studies could not be fruitfully performed by non-specialists - even though the software tools were often easily accessed and easily used by those unfamiliar with computational chemistry. It was unanimously accepted that the performance of *in-silico* studies in CHM research necessitated close interaction between computational chemists and CHM experimentalists.

1.3 Discussion on evaluation criteria for CHM databases

In the lead up to the consensus meeting, a draft set of criteria for evaluation of CHM databases was prepared by Barlow and was circulated for consideration by the WP4 *in-silico* work-group; these criteria (presented in Table 1) served to provide a focus for discussions at the consensus meeting.

Table 1 – Draft list of quality evaluation criteria for CHM databases

Evaluation Criteria	***Description of criteria***	***Score***
Availability & Cost	Is the database publicly available (either free or at a discount for academics)	**100**
Platform	Is the database implemented under Windows or Linux or accessed via the internet?	**10**
Data	Does the database include chemical and bibliographic information on the plants' chemical constituents?	**10**
Search facilities	Can the database be searched with queries covering chemical, botanical, and/or TCM usage?	**10**
Results	Can the hits from searches be saved to output file(s)?	**10**

E-mails received from the *in-silico* work-group members prior to the March meeting noted the following deficiencies in the draft CHM database evaluation criteria:

1. Compared to Availability and Cost the other scores are rather low. The first criterion is clearly more important, but higher scores for some of the other criteria would help to distinguish a 'good' publicly available database from a 'bad' publicly available database with greater ease.

2. The Availability and Cost should be maintained as a criterion, given that sources which are not publicly available will not be of interest to most people. However, databases that are not publicly available will probably be of interest to researchers who are perhaps on the lookout for unusual information, and might then wish to contact the database authors directly.

3. As far as 'platform' is concerned, it would be useful to include this, but to make it clear whether there is a website, or a CD/DVD, or both. For intensive use, a CD/DVD is much preferable to online access.

4. In 'Results' it would be preferable to distinguish between databases where users can output a list or batch file of chemical structures from a search, and those where this is not possible. It's often the case, particularly with commercial databases, that structures have to be downloaded individually, and this must be regarded as a significant limitation/deficiency.

5. The scores under the criterion of Availability and Cost should distinguish between those databases that are i) free, ii) available at academic discount, and iii) available only at full cost, with suggested scores of. i) 100, ii) 75, and iii) 25, respectively

6. The scoring should take into account the nature of the available chemical data, such as whether the structure is 2D or 3D, whether the compound stereochemistry is unequivocally defined or different enantiomers should be generated and tested, etc.

Further discussions held at the London consensus meeting, raised the following additional issues and considerations:

1. Databases that are *not* publicly available probably should not be considered here. A lack of availability was thus proposed as an Exclusion criterion.

2. Platform implementation is probably not that important as regards CHM databases and could be removed as an evaluation criterion.

3. The criteria should permit discrimination between those databases that hold only 2D chemical structure data, those that hold 3D data on single conformers, and those that hold 3D data on multiple conformers.

4. The criteria should permit discrimination between databases that do and do not cater for batch output of chemical structures.

5. The criteria should permit discrimination between databases that are static and those that are live and curated.

6. The criteria should reward databases that are reported in a peer-reviewed scientific publication.

7. Databases should be scored so as to reward the range of data included, information on TCM usage, toxicology data, known or predicted target(s), *etc*.

8. Databases should be differentially scored according to the complexity they cater for in terms of search fields.

1.4 Consensus criteria for evaluation of CHM databases

Taking due consideration of all points raised in e-communications received prior to the London meeting, together with those raised in discussions at the meeting, a final set of quality evaluation criteria were formulated as presented in Table 2. With these criteria, it was thus proposed that the available CHM databases should be scored on the basis of their content (criteria C1-C11), the versatility of Boolean searches catered for (criteria C12-C14), the nature of the allowed chemical structure output (criteria C15-C16), and the availability of associated reports in peer-reviewed publications (criteria C17). The maximum score achievable is 50.

Table 2 – Quality evaluation criteria for CHM databases

Evaluation Criteria	*Description of criteria*	*Score*
C1	Contains information on CHM usage?	**2**
C2	Contains botanical information of plants used in CHM?	**2**
C3	Contains toxicological data on CHM phytochemicals?	**2**
C4	Contains (known) biological activity data on CHM phytochemicals?	**2**
C5	Contains (predicted) biological activity data on CHM phytochemicals?	**2**
C6	Contains chemical structure data on CHM phytochemicals?	**2**
C7	Contains 2D structures of CHM phytochemicals?	**2**
C8	Contains 3D structures of CHM phytochemicals?	**2**
C9	Contains data on the molecular mechanism(s) of CHM phytochemicals?	**2**
C10	Data on >1,000 unique CHM phytochemicals	**2**
C11	Database is live, curated and updated?	**2**
C12	Database can be queried by chemical structure?	**2**
C13	Searches with ≥ 2 search fields possible?	**2**
C14	Searches can be performed using combinations of search fields?	**2**
C15	Chemical structure(s) can be output in 1D (as SMILES strings)?	**1**
	Chemical structure(s) can be output in 2D?	**2**
	Chemical structure(s) can be output in 3D (as single conformers)?	**4**
	Chemical structure(s) can be output in 3D (as conformer libraries)?	**8**
C16	Facility for (batch) output of multiple chemical structures	**5**
C17	Database is reported in an article in a peer reviewed journal	**2**

1.5 Consensus criteria for evaluation of CHM databases

The quality evaluation criteria presented in Table 2, were subsequently applied in evaluation of the existing CHM databases (as catalogued in Tables 1 and 2 under D4.11). The consensus scores for these various databases are presented in Table 3.

Table 3 – Consensus evaluation scores for CHM databases

Database*	Content	Source†	URL	Overall Score
TCM Assistant	TCM herbs, herbal formulas, diseases and patent prescriptions. No structures	-	http://www.tcmassistant.com	**10**
Dictionary of Chinese Herbs	TCM herbal formulas, specificity, toxicity and side effects. No structures.	-	http://Alternativehealing.org/Chinese_herbs_dictionary.htm	**10**
Dictionary of Natural Products (DNP)	Major source of chemical information on natural products, including some biological sources, and pharmacological and toxicological data. Full set of structures.	-	http://dnp.chemnetbase.com	**21**
China Natural Products Database (CNPD)	Information on Chinese natural products including >40,000 structures. Full set of structures.	Shen *et al.*[2]	http://www.neotrident.com	**27**
3D Structure Database of Components from Chinese Herbs	3D structures (>10,000) from Chinese herbs (>2,000), with descriptors and data on clinical uses. Full set of structures.	Qiao *et al.*[3]	-	**n/d**
Traditional Chinese Medicines Database (TCMD)‡	Information and structures for >10,000 compounds from >4,500 species. Full set of structures.	Yan *et al.*[4]	http://www.cambridgesoft.com	**n/d**

TCM Knowledge Based Grid	TCM herb database, literature database, traditional Tibetan herb database. No structures.	-	http://www.cintcm.com	**8**
Chinese herbal constituents database (CHCD) and Bioactive plant compounds database (BPCD)	Information and structures for >13,000 constituents of ≈300 commonly used herbs. >2,500 compounds active against ≈80 targets. Full set of structures.	Ehrman *et al.*[5]	-	**n/d**
Dr. Duke's Phytochemical and Ethnobotanical Databases	Information on phytochemicals from >1000 plants, including Chinese herbs. No structures.	Duke[6]	http://www.ars-grin.gov/duke/	**14**
PhytochemDB	Chemical composition of 1,278 taxa (>19,000 constituents), including Chinese herbs. No structures.	-	http://ukcrop.net/perl/ace/search/PhytochemDB	**8**
Ethnopharmacological database (GPNDB™)	100,000 natural products (3D structures), biological activities, ethnopharmacological data. In-house database of Greenpharma S.A.	Do & Bernard[7]	http://www.greenpharma.com	**n/d**
Comprehensive Herbal Medicine Information System for Cancer (CHMIS-C)	Integrated information on cancer molecular targets, Chinese herbal recipes and phytochemical constituents. Some structures.	Fang *et al.*[8]	http://sw16.im.med.umich.edu/chmis-c/	**37**
Traditional Chinese Medicine Information Database (TCM-ID)	Information on 1197 formulas, 1098 herbs and 9852 constituents in relation to TCM diagnosis and prescription. Some structures.	Chen *et al.*[9]	http://tcm.cz3.nus.edu.sg/group/tcm-id/tcmid.asp	**22**

Therapeutic Target Database (TTD)	Information on 1,894 targets, 5,028 drugs, diseases, and pathways.	-	http://xin.cz3.nus.edu.sg/group/ttd/ttd.asp	**24**
Potential Drug Target Database (PDTD)	Information on 830 targets, protein and active site structures, biological functions, diseases and pathways.	Fang *et al.*[10]	http://www.dddc.ac.cn/pdtd/	**6**
Traditional Chinese Medicine Database System	Bibliographic database (TCMLARS), and Chinese herb database (TCDBASE) in addition to other data.	Chen[11]	http://www.cintcm.com	**8**
TCMGeneDIT	Information on relations between TCM and gene regulation, protein-protein interactions and biological pathways.		http://tcm.lifescience.ntu.edu.tw/	**18**
TCM Database@Taiwan	Chinese medicine database that contains 3-D structural information of TCM constituents - ready for molecular docking simulation (database currently holds 37,170 (32,364 non-duplicate) TCM compounds from 352 TCM)		http://tcm.cmu.edu.tw/review.php?menuid=3	**30**

† References cited are those listed numerically in WP4 deliverable D4.11

* Databases not publicly available (and which could not, therefore, be evaluated) are highlighted, and their overall scores shown as n/d

‡ The TCMD database, originally marketed by CambridgeSoft, is no longer available, and the company has withdrawn support.

It may be noted (Table 3) that 4 of the 18 available CHM databases could not be scored because they were not available for evaluation. Of the remaining 14 databases, those with the highest scores, and the ones therefore that are considered to more useful for *in-silico* studies of CHM, include: the Chinese Natural Products Database (score =2 27), the TCM Database@Taiwan (score = 30), and the Comprehensive Herbal Medicines Information System for Cancer (score = 37). The former two databases are quite general, while the latter is focussed on compounds of relevance to cancer research.

1.6 Discussion on evaluation criteria for *in-silico* tools

In the lead up to the consensus meeting, a draft set of criteria for evaluation of *in-silico* tools for use in CHM research was prepared by Barlow and was circulated for consideration by the WP4 *in-silico* work-group; these criteria (presented in Table 4) served to provide a focus for discussions at the March consensus meeting.

Table 4 – Draft list of quality evaluation criteria for *in-silico* analytical tools

Evaluation Criteria	***Description of criteria***	***Score***
Availability & Cost	Is the software publicly available (either free or at a discount for academics)?	**100**
Platform	Does the software run under Windows, or Linux or can it be accessed via the internet?	**10**
Methods - Algorithms	Do the algorithms employed have proven utility in other (relevant) problem areas and/or have validated utility in TCM studies	**5**
Methods - Statistics	Are the output(s) provided by the software qualified by some measure of statistical reliability/certainty	**5**
Methods – Ease of use	Is the software easy for non-specialists to use, and/or are there (informative, easy to digest) manuals/handbooks available to guide use?	**5**
Results	Are the output(s) generated within a convenient timeframe (and is the software amenable, therefore, to high throughput use)?	**5**

E-mails received from the *in-silico* work-group members prior to the March meeting noted the following deficiencies in the draft *in-silico* tools evaluation criteria:

1. As considered for CHM databases, the scores under the criterion of Availability and Cost should distinguish between those tools that are i) free, ii) available at academic discount, and iii) available only at full cost, with suggested scores of. i) 100, ii) 75, and iii) 25, respectively.

2. Platform is relevant, since high throughput screening approaches require a very efficient use of RAM and a stable operating system, and UNIX/MacOSX/Linux operating systems are generally more suitable in this regard than Windows. In addition, it was noted that the utility of web–based resources is critically dependent on their availability. A modified scoring scheme was proposed as: i) UNIX/Linux (MacOSX included) 10, ii) Windows 8, iii) Web/Others 5.

3. It is not necessary for the algorithms to have been validated specifically for use in CHM research; it would suffice that they have been validated on other ligand sets that exhibit similar chemical diversity.

4. *In-silico* tools should ideally ***not*** be used by non-specialists, since almost everyone can obtain an output from a computational program, but - if not carefully produced/checked the output obtained could be meaningless. That said, manuals/handbooks should be available and should be exhaustive.

1.7 Results from the questionnaires and virtual discussion group on use of functional genomic techniques for *in-vitro* CHM research

Additional input in consideration of good practice in the use of *in-silico* tools in CHM research was provided by the responses returned by WP4 members and the *in-silico* work-group after circulation of the questionnaire prepared under WP4 D4.14.

The questionnaire was returned by 10 out of the 11 WP4 members and also by 3 of the 5 WP4 *in-silico* work-group. Several of the WP4 respondents, however, did not feel they could answer questions 8 and 9 on *in-silico* research due to lack of expertise.

Full sets of answers to all nine questionnaire questions from returning WP4 members is included in Annex 1 of WP4 D4.14; all answers received relating to questions 8 and 9, which pertain to *in-silico* studies, are reproduced below.

1.7.1 QUESTION 8: How can *in-silico* tools be best applied to CHM research?

Member 1: Chemo-genomics may be the single most important objective. However, even in the event that many targets of many compounds can be predicted with reasonable accuracy, there remains the question of how best to use those data. This requires input from bio-informatics (gene/disease relationships), and also TCM. Bridging the gulf between medical paradigms is something to which data mining is well suited.

Member 2: The *in silico* tools are very useful in the first steps of pharmacological research. Computational approaches allow us to evaluate a lot of molecules on several targets with minimal economic effort. The positive hits can subsequently be investigated through *in vitro* and/or *in vivo* approaches, which are more expensive and associated with lower throughput.

Member 3: *In silico* tools and more in general bio-informatics/computer biology are just essential for functional genomics application, both for multiple data analysis and systems modelling. Without their analytical power would be impossible to approach such complex experimental tools.

Member 4: For evaluation and interpretation of the results obtained from functional genomics approaches.

Member 5: In my opinion, they can be used as predictive tools for pointing out or directing molecular targets or mechanisms involved in TCM actions. They might narrow the amount of targets to be checked. Additionally, they are useful for interpretation of 'omics results.

Member 6: Next to the mandatory use of bio-informatics in data evaluation/interpretation, in silico tools such as pharmacophore modeling and in silico screening may be used to get first ideas about biological targets (If there is any knowledge about the chemical structure of single compounds available)

Member 7: Built on ever –strengthening bio-informatics technologies, *in-silico* tools could be useful in finding patterns, pathways, effector/mediator clusters and help forge new hypotheses and guide future research.

Member 8: *In silico* tools are best applied in CHM research as a means: to seek out potential mechanisms of action of their constituents; to identify putative new leads for drugs; and to summarise and/or visualize the complex patterns embedded within the output generated through associated 'omics studies. Given the complexity of CHM, and the ensuing difficulties in performance of experimental studies, *in-silico* studies clearly offer an economical and efficient

way of exploring the problem landscape and thus helping to define suitable hypotheses for testing in the laboratory.

1.7.2 QUESTION 9: Are *in-silico* approaches readily available and what is needed to potentiate their application in TCM research

Member 1: Most of the tools that have been developed so far either revolve around data (generally fairly simple inventories of relevant compounds) and docking (eg. TarFisDock and iScreening). iScreening will apparently dock 20,000 compounds from CHM against any PDB receptor which the user uploads, and send you a list of the best 200 compounds. The main difficulties are that many of the databases are difficult to use or access. The efficacy of the current docking tools is also questionable (in terms of speed, reliability etc.). Though they are, for the most part, easy to operate, they are not suitable for those with no background in virtual screening, and the results are intended for individual use (though there is probably no reason why they can't be collected by the provider in the form of a database). However, as yet, there is no centralized repository which combines information both on compounds and targets, traditional uses and so on, which can be easily accessed by users from a wide variety of backgrounds. So far, it is therefore the case that *in silico* approaches are at a fairly low level of organization/sophistication. Raw data and docking results for a limited number of compounds and/or targets comprise the first stage. There are further layers of analysis that need to be undertaken before a more complete picture begins to emerge. As mentioned above, these include bio-informatics and TCM. Many of the appropriate tools are now in place, though refinements to most of them will continue to be made. That may not be as important as the application of *different* methods of virtual screening to compound databases (eg. ligand-based, structure-based pharmacophores and docking), and the use of results based on consensus (which should become progressively easier).

Member 2: The use/implementation of *in silico* approaches require specific expertises in computer science and computational chemistry, since the risk of producing biased results is not null. An efficient application of in silico approaches requires the chemical characterization of putative active compounds and their organization in chemical/structural databases.

Member 3: what is really needed is a better crosstalk between the in vitro and the in silico experts as well as an improved academic curriculum for in vitro scientists in order for them to better understand the language and the potentialities of comuter science for their research

Member 4: In silico approaches and tools that facilitate and complement molecular biology studies are very dynamic research area that quickly develops. Direct interaction with experts from that area is still is the best option to maximise the value of the obtained in silico

predictions.

Member 5: Today, In silico approaches are mostly available for most of the scientist, although good use and successful results from their use required a strong knowledge of them. On the other hand, to get results from in silico approaches, the scientifically question to be solved should be accurately formulated to avoid getting tons of non-usable or non-interpretable information. Finally, in silico results must been confirmed by other molecular and cellular approaches.

Member 6: Here strong collaborations are needed. And collaborations are only readily available when the scientific field attracts expert in this field.

Member 7: I think in silico approach is been used all the time in such an Information Technology era. For example, analysis of conventional and “omics” data are routinely carried out using computerised tools worldwide. There are also many databases with speclialised analysing powers, which are developed in the past years, but they are usually poorly maintained, updated and their accessibility is generally poor. Meanwhile there is much duplication in efforts, re-spinning the wheels and wasting limited resources. Sustainable support from governments, charities and the industrial community is needed for establishing authoritative and well-maintained systems; reliable multiple language translation is another on our wish list; open access or easy access will also make a great difference.

Member 8: There are numerous *in silico* tools available for use in CHM research, and also a significant number of databases of relevance in such studies. The software tools are not always readily applied by those who are not specialists in the area, however, and the outputs generated are best interpreted with a caution that comes from expertise. Here, therefore, as with bio-informatics and chemo-informatics generally, it is essential that the (*in vitro*) experimentalists should collaborate closely with the computational chemists in order that such studies furnish *meaningful* data.

1.8 Conclusions as regards the use of *in-silico* tools in CHM research

1. *In-silico* tools are best applied in CHM research through close collaboration between computational specialists and CHM experimentalists.
2. A scoring scheme for evaluation of *in-silico* tools seems unnecessary.
3. The usefulness and choice of *in-silico* screening tool is dependent on the reason for its use and/or the nature of the output to be generated.
4. *In silico* tools are best applied in CHM research as a means: to seek out potential mechanisms of action of their constituents; to identify putative new leads for drugs; and to summarise and/or visualize the complex patterns embedded within the output generated through associated 'omics studies. Given the complexity of CHM, and the ensuing difficulties in performance of experimental studies, *in-silico* studies clearly offer an economical and efficient way of exploring the problem landscape and thus helping to define suitable hypotheses for subsequent testing in (*in-vitro* and/or *in-vivo*) laboratory studies.