Advances in Intelligent Systems and Computing 294

Julio Sáez-Rodríguez
Miguel P. Rocha
Florentino Fdez-Riverola
Juan F. De Paz Santana *Editors*

8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)

Springer

Advances in Intelligent Systems and Computing

Volume 294

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/11156

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

Julio Sáez-Rodríguez · Miguel P. Rocha
Florentino Fdez-Riverola · Juan F. De Paz Santana
Editors

8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)

Springer

Editors
Julio Sáez-Rodríguez
European Bioinformatics Institute
Hinxton
United Kingdom

Miguel P. Rocha
University of Minho
Braga
Portugal

Florentino Fdez-Riverola
Department of Informatics
University of Vigo
Ourense
Spain

Juan F. De Paz Santana
Department of Computing Science
and control
University of Salamanca
Salamanca
Spain

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-07580-8 ISBN 978-3-319-07581-5 (eBook)
DOI 10.1007/978-3-319-07581-5
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014939943

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Biological and biomedical research are increasingly driven by experimental techniques that challenge our ability to analyse, process and extract meaningful knowledge from the underlying data. The impressive capabilities of next generation sequencing technologies, together with novel and ever evolving distinct types of omics data technologies, have put an increasingly complex set of challenges for the growing fields of Bioinformatics and Computational Biology. To address the multiple related tasks, for instance in biological modeling, there is the need to, more than ever, create multidisciplinary networks of collaborators, spanning computer scientists, mathematicians, biologists, doctors and many others.

The International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB) is an annual international meeting dedicated to emerging and challenging applied research in Bioinformatics and Computational Biology. Building on the success of previous events, the 8th edition of PACBB Conference will be held on 4–6 June 2014 in the University of Salamanca, Spain. In this occasion, special issues will be published by the Journal of Integrative Bioinformatics, the Journal of Computer Methods and Programs in Biomedicine and the Current Bioinformatics journal covering extended versions of selected articles.

This volume gathers the accepted contributions for the 8th edition of the PACBB Conference after being reviewed by different reviewers, from an international committee composed of 72 members from 15 countries. PACBB'14 technical program includes 34 papers from about 16 countries of origin, spanning many different sub-fields in Bioinformatics and Computational Biology.

Therefore, this event will strongly promote the interaction of researchers from diverse fields and distinct international research groups. The scientific content will be challenging and will promote the improvement of the valuable work that is being carried out by the participants. Also, it will promote the education of young scientists, in a post-graduate level, in an interdisciplinary field.

We would like to thank all the contributing authors and sponsors (Telefónica Digital, Indra, Ingeniería de Software Avanzado S.A, IBM, JCyL, IEEE Systems Man and Cybernetics Society Spain, AEPIA Asociación Española para la Inteligencia Artificial, APPIA Associação Portuguesa Para a Inteligência Artificial, CNRS Centre national

de la recherche scientifique), AI*IA, as well as the members of the Program Committee and the Organizing Committee for their hard and highly valuable work and support. Their effort has helped to contribute to the success of the PACBB'14 event. PACBB'14 wouldn't exist without your assistance. This symposium is organized by the Bioinformatics, Intelligent System and Educational Technology Research Group (http://bisite.usal.es/) of the University of Salamanca and the Next Generation Computer System Group (http://sing.ei.uvigo.es/) of the University of Vigo.

Julio Sáez-Rodríguez
Miguel P. Rocha
PACBB'14 Programme Co-chairs

Florentino Fdez-Riverola
Juan F. De Paz Santana
PACBB'14 Organizing Co-chairs

Organization

General Co-chairs

Florentino Fdez-Riverola	University of Vigo, Spain
Juan F. De Paz	University of Salamanca, Spain
Julio Sáez-Rodríguez	European Bioinformatics Institute, United Kingdom
Miguel Rocha	University of Minho, Portugal

Program Committee

Alicia Troncoso	University Pablo de Olavide, Spain
Amparo Alonso	University of A Coruña, Spain
Ana Cristina Braga	University of Minho, Portugal
Anália Lourenço	University of Vigo, Spain
Armando Pinho	Universty of Aveiro, Portugal
Caludine Chaouiya	Gulbenkian Institute, Portugal
Camilo Lopez	Universidad Nacional de Colombia, Colombia
Carlos A.C. Bastos	University of Aveiro, Portugal
Daniel Glez-Peña	University of Vigo, Spain
Daniela Correia	CEB, University of Minho, Portugal
David Hoksza	Charles University in Prague, Czech Republic
Eva Lorenzo	University of Vigo, Spain
Fernanda Correia Barbosa	DETI/IEETA, University of Aveiro, Portugal
Fernando Díaz-Gómez	University of Valladolid, Spain
Fidel Cacheda	University of A Coruña -, Spain
Florencio Pazos	CNB, Spanish Council for Scientific Research, Spain
Francisco Torres-Avilés	Universidad de Santiago de Chile, Chile
Frank Klawonn-Ostafilia	University of Applied Sciences, Wolfenbuettel, Germany
Gonzalo Gómez-López	UBio/CNIO, Spanish National Cancer Research Centre, Spain

Gustavo Isaza	Universidad de Caldas, Colombia
Hagit Shatkay	University of Delaware, USA
Heri Ramampiaro	Norwegian University of Science and Technology, Norway
Hugo López-Fernández	University of Vigo, Spain
Hugo Miguel Santos	Universidade Nova de Lisboa, Portugal
Isabel C. Rocha	IBB/CEB, University of Minho, Portugal
Jiri Novak	Charles University in Prague, Czech Republic
João Rodrigues	University of Aveiro, Portugal
Joel P. Arrais	DEI/CISUC, University of Coimbra, Portugal
Jorge Ramirez	Universidad Nacional de Colombia, Colombia
Jorge Vieira	Institute for Molecular and Cell Biology, Portugal
José Antonio Castellanos Garzón	University of Valladolid, Spain
Jose Ignacio Requeno	University of Zaragoza, Spain
José Luis Capelo	Universidade Nova de Lisboa, Portugal
José Luis Oliveira	Universty of Aveiro, Portugal
José Manuel Colom	University of Zaragoza, Spain
Juan Antonio García Ranea	University of Malaga, Spain
Julio R. Banga	IIM, Spanish Council for Scientific Research, Spain
Liliana Lopez-Kleine	Universidad Nacional de Colombia, Colombia
Loris Nanni	University of Bologna, Italy
Lourdes Borrajo	University of Vigo, Spain
Luis F. Castillo	Universidad de Caldas, Colombia
Luis Figueiredo	European Bioinformatics Institute, United Kingdom
Luis M. Rocha	Indiana University, USA
M Alamgir Hossain	Northumbria University at Newcastle, United Kingdom
Mª Araceli Sanchís de Miguel	University Carlos III of Madrid, Spain
Manuel Álvarez Díaz	University of A, Spain
Miguel Reboiro	University of Vigo, Spain
Mohammad Abdullah Al-Mamun	Northumbria University, United Kingdom
Mohd Saberi Mohamad	Universiti Teknologi Malaysia, Malaysia
Monica Borda	University of Cluj-Napoca, Romania
Narmer Galeano	Cenicafé, Colombia
Nuno Fonseca	CRACS/INESC, Porto, Portugal
Nuria Medina Medina	CITIC, University of Granada, Spain
Pierpaolo Vittorini	University of L'Aquila, Italy
Reyes Pavón	University of Vigo, Spain
Rita Ascenso	Polytecnic Institute of Leiria, Portugal
Rosalía Laza	University of Vigo, Spain

Rubén López-Cortés	Universidade Nova de Lisboa, Portugal
Rui Brito	University of Coimbra, Portugal
Rui C. Mendes	CCTC, University of Minho, Portugal
Rui Camacho	LIAAD/FEUP, Universty of Porto, Portugal
Rui Rijo	Polytecnic Institute of Leiria, Portugal
Sara C. Madeira	IST/INESC ID, Lisbon, Portugal
Sara P. Garcia	University of Aveiro, Portugal
Sérgio Deusdado	Polytecnic Institute of Bragança, Portugal
Sergio Matos	DETI/IEETA, University of Aveiro, Portugal
Silas Vilas Boias	Univerity of Auckland, New Zealand
Slim Hammadi	Ecole Centrale de Lille, France
Thierry Lecroq	Univeristy of Rouen, France
Tiago Resende	CEB, University of Minho, Portugal
Vera Afreixo	University of Aveiro, Portugal

Organising Committee

Juan M. Corchado	University of Salamanca, Spain
Javier Bajo	Polytechnic University of Madrid, Spain
Juan F. de Paz	University of Salamanca, Spain
Sara Rodríguez	University of Salamanca, Spain
Dante I. Tapia	University of Salamanca, Spain
Fernando de la Prieta Pintado	University of Salamanca, Spain
Davinia Carolina Zato Domínguez	University of Salamanca, Spain
Gabriel Villarrubia González	University of Salamanca, Spain
Alejandro Sánchez Yuste	University of Salamanca, Spain
Antonio Juan Sánchez Martín	University of Salamanca, Spain
Cristian I. Pinzón	University of Salamanca, Spain
Rosa Cano	University of Salamanca, Spain
Emilio S. Corchado	University of Salamanca, Spain
Eugenio Aguirre	University of Granada, Spain
Manuel P. Rubio	University of Salamanca, Spain
Belén Pérez Lancho	University of Salamanca, Spain
Angélica González Arrieta	University of Salamanca, Spain
Vivian F. López	University of Salamanca, Spain
Ana de Luís	University of Salamanca, Spain
Ana B. Gil	University of Salamanca, Spain
Mª Dolores Muñoz Vicente	University of Salamanca, Spain
Jesús García Herrero	University Carlos III of Madrid, Spain

Contents

Applications

Data Analysis and Mining

Proteins

Sequence Analysis

Systems Biology

Text Mining

Agent-Based Model for Phenotypic Prediction Using Genomic and Environmental Data

Sebastien Alameda[1], Carole Bernon[1], and Jean-Pierre Mano[2]

[1] Universite Paul Sabatier, Toulouse, France
[2] UPETEC, Toulouse, France

Abstract. One of the means to increase in-field crop yields is the use of software tools to predict future yield values using past in-field trials and plant genetics. The traditional, statistics-based approaches lack environmental data integration and are very sensitive to missing and/or noisy data. In this paper, we show how using a cooperative, adaptive Multi-Agent System can overcome the drawbacks of such algorithms. The system resolves the problem in an iterative way by a cooperation between the constraints, modelled as agents. Results show a good convergence of the algorithm. Complete tests to validate the provided solution quality are still in progress.

Keywords: Multi-Agent System, Adaptation, Self-organization, Phenotypic Prediction.

1 Introduction

Constant growth in global population, hence cereal consumption, increases the pressure on food processing industries to meet this increasing demand[1]. In this context, human and commercial necessity to produce more and more cereals implies the use of industrial processes that guarantee higher in-field yields. Amongst these processes, genomic breeding is a widely used set of techniques encompassing mathematical and software tools able to predict a crop yield based on genetics[2]. These tools, currently statistics-based, can be used to improve yield by choosing plant varieties with higher genetic potentials. The statistical methods traditionally used for these purposes lack the integration of environmental conditions in the predicting variables. Therefore, they have yet been unable to predict the yield variability of a crop depending on the weather - and other environmental parameters, such as the ground quality - it is exposed to.

To overcome this lack, we aim at building a system able to predict a yield value, depending on experimental conditions, by using cooperative, adaptive, self-organizing agent-based techniques. This system ought to be able to use raw data without any preprocessing. The experiments run on this system use data provided by seed companies, extracted from in-field maize experiments, which are both noisy and sparse. To validate this system, leave-one-out test cases will be executed to check its convergence.

J. Sáez-Rodríguez et al. (eds.), *8th International Conference on Practical Appl. of Comput. Biol. & Bioinform. (PACBB 2014)*, Advances in Intelligent Systems and Computing 294,
DOI: 10.1007/978-3-319-07581-5_1,

2 Problem Expression

2.1 Original Problem in Genomic Breeding

The problem is to predict the γ_i phenotype of an individual i ($i = 1..n$) knowing a $1 * p$ vector x_i of SNP genotypes on this individual. It is generally assumed that

$$\gamma_i = g(x_i) + e_i \tag{1}$$

with g being a function relating genotypes to phenotypes and e_i an error term to be minimized. The γ_i value found once the g function is computed is called the Genomic Estimated Breeding Value [3].

To find the actual value of the g function, i.e. to minimize the e_i terms, various methods can be used. Without drifting into too much detail, Random Regression Best Linear Unbiased Prediction (RR-BLUP) [4] offers good results in the context of biparental crosses, which is the case in maize breeding. This method, and the others used in plant breeding, have in common the goal to minimize, as said before, the error term and to find an accurate, global, expression of g.

Those global approaches pose the problem of the quality of the data involved in the predictions. For example it has been shown that the marker density - the size of the x_i vector related to the genome size of the considered species - needs to scale with population size and that the choice of the samples are of great importance in the accuracy of the results [2].

Furthermore, the accuracy of the prediction given by those models depends heavily on trait heritability. The more a trait is heritable, the more accurate the prediction [5]. This lack of accuracy in low-heritability traits may be explained by the influence of environmental parameters on those traits and by genomic-environmental interactions and brings the need for another problem expression able to integrate environmental data.

2.2 Problem Specification

The problem this paper addresses is the prediction of the γ yield value of a maize crop given a set x_i of n constraints on various genetical and environmental traits. The equation (1) becomes:

$$\gamma = g(x_i) + e \tag{2}$$

with g being a continuous function and e being the error term.

The assumed continuity property of g allows a local, exploratory search of the solution. In other terms, it removes the need of finding a global, search space wide definition for g. The means we offer to find a solution is to iteratively fetch relevant data on previously measured in-field tests from a database. To be deemed "relevant", a datum must match the constraints expressed by the x_i vector.

As discussed above, the relevant data $\{D_i\}$ are extracted from a database of past in-field trials on the basis of the constraints defined by the x_i parameters.

As the database typically holds more than a million of such data and can theoretically contain much more, for scalability purposes only a few of them is loaded in the memory at each iteration. Each datum D_i that constitutes the dataset is itself a set encompassing, for an observed phenotype, all phenotypic, environmental and genomic data related to this phenotype. In particular, the datum D_i holds a γ_i value for the phenotypic trait that is the goal of the prediction.

One of the challenges that the system must address is to cooperatively decide which constraints should be individually released or tightened, i.e. the tolerance to add to each constraint, in order to reach a satisfactory solution. Since a solution is defined as a dataset $\{D_i\}$, in the ideal case, all γ_i would be equal to one another (consistent solution) and the data set would contain a large number of data (trustworthy solution). Such a solution is deemed "satisfactory" when the system cooperatively decides that it cannot be improved anymore.

The solution satisfaction can then be expressed as a f_a function, aggregation of two functions:

- A function f_q that evaluates the quality of the solution as the range taken by the predicted values $\{\gamma_i\}$. The lower this range, the lower the value of $f_q(\{D_i\})$.
- A function f_t that evaluates the trust given to the solution provided. The more data D_i are implied in the solution, the lower the value of $f_t(\{D_i\})$.

With this definition, the goal of the prediction system is expressed as providing a solution $\{D_i\}$ as close as possible to the absolute minimum of f_a.

Linking back to the equation (2), $g(x_i)$ may then be defined as the average value of the $\{\gamma_i\}$ and e as a term bounded by the range of $\{\gamma_i\}$.

3 Solving Process

Agents are defined as autonomous entities able to perceive, make decisions and act upon their environment [6]. A system of those interconnected software agents is able to solve complex problems. The system used in order to solve this problem is based on the AMAS (Adaptive Multi-Agent System) theory [7], which provides a framework to create self-organizing, adaptive and cooperative software. The agents in the system, by modifying their local properties (adaptive) and the interactions between them (self-organizing), modify also the global function of the system.

3.1 The System and Its Environment

The AMAS considered here contains two different kinds of agents:

- n Constraint Agents, in charge of tightening or releasing the constraints defined in section 2.2. Each agent is responsible for one constraint. Each agent's goal is to minimize its estimation of the f_a function, calculated on the only basis of this agent's actions.

– A Problem Agent, in charge of evaluating the solution provided by the Constraint Agents and giving them a hint on the future actions they have to take in order to make the solution more satisfactory. Its goal is to minimize the actual f_a function.

3.2 Iterative Process

The resolution is iterative and the Fig.1 illustrates the way the system functions.

At each step, the Problem Agent (1) receives a data set $\{D_i\}$ and evaluates both values of $f_q(\{D_i\})$ and $f_t(\{D_i\})$. Every Constraint Agent (2) has three possible actions: tightening, releasing or leaving as is the constraint it is related to.

In order to decide amongst its possible actions, each Constraint Agent evaluates its influence on the solution quality f_q and the solution trust f_t by simulating the world state if it were to execute one or the other of its possible actions. Depending on this simulated state of the world, the agent chooses the most cooperative action to perform, that is the action that improves the value of the criterion the agent has the greatest influence on. This calculated influence gives the agent a hint on whether it should maintain as is, tighten or release the constraint it is related to.

The current restriction state of the constraints are aggregated (3) and used as a filter to find a new dataset $\{D_i\}$. This dataset consists of previously found data matching the new constraints and newly found data, also matching these new constraints, from the database (4). This way, the system simply ignores the missing data by including in the datasets only the existing, relevant data.

At each step, Each datum D_i in the database can be in one of these three states:

– Active: the datum is loaded into memory and, at each resolution step, gives a predicted value γ_i.
– Inactive: The datum was loaded into memory once but does not provide predicted values, as it does not match one of the current constraints.
– Existing: The datum exists in the database but has not currently been loaded into memory.

This model allows an iterative enrichment of the data pool. As the constraints become more precise regarding the problem to be solved, the Inactive + Active pool size tends to remain constant due to the fact that every datum matching the constraints has already been loaded into memory and no more data are loaded from the Existing data pool.

3.3 Convergence Measurement

The resolution ends when the dataset $\{D_i\}$ provided at the end of each resolution step is definitely stable. To guarantee this stability, two conditions must be met:

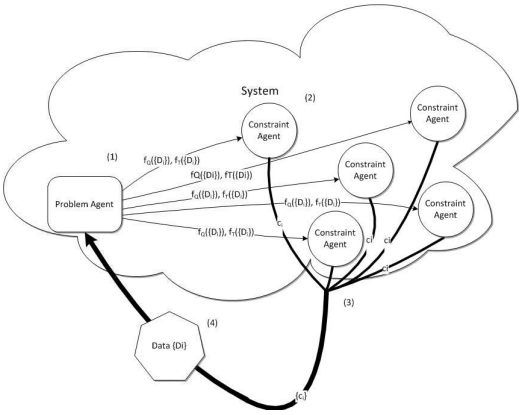

Fig. 1. A view of the system architecture exhibiting the information flow between the agents

- Every Constraint Agent estimates that the optimal (from its own point of view) action to take is to not modify its value.
- The Active + Inactive dataset size is stable, i.e. no more data are recruited from the database.

In those conditions, the system has reached a fixed point and the convergence process is complete. At this point, the data matching the constraints constitute the solution provided to the user.

4 Experiments and Results

As seen above, the convergence is characterized by the stability of the constraints and the stability of the Inactive + Active dataset size. The primary objective of the following experiments is to exhibit those two convergence conditions. The other objectives are to show that the convergence speed and the quantity of data used make this AMAS solution suitable for real-life use.

The experimental protocol set up is the random choice of several leave-one-out test cases. The data used are real-world in-field maize data, provided by seed companies that are partners of this research project.

4.1 Data Characterization

These data include:

- 300,000 maize individuals with their pedigree and/or genomic data;
- 30,000,000 yield and other phenotypical data of in-field trials in the past years for these individuals;
- 150,000 environmental (meteorological and pedological) data for these trials.

Those data are essentially sparse with respect to the various dependent variables in this problem. Indeed, the phenotypical measurements result from the interaction of a given maize individual, identified by its genomic data, and a specific environment, which can be uniquely determined by a given location and year, in which interfere the various environmental data specified above. If one considers for instance that these data measurements are arranged in a rectangular matrix, with individuals per rows and environments per columns, then the resulting matrix will be extremely sparse, i.e. with a high ratio of zero entries corresponding to unobserved data. This sparsity aspect is intrinsic to the problem, simply because it is infeasible to grow every year in every location all the existing maize individuals. With respect to the database considered here, in the case of the yield values (which is one of the most frequently collected data), the ratio of the number of measured values to the total number of entries in this matrix is less than 0.7 percent. In [8], the authors recall either techniques that try to input the missing data in some way, or methods that are designed to work without those missing input values, the first ones being sensitive to the ratio of observed to missing data, and the latter presenting some risk of overfitting. The AMAS method we consider here belongs to the second class of methods, and present the additional advantage that it does not suffer from overfitting issues, since the method itself aims at selecting a much denser subset of values that are relevant for a given problem.

As the data are provided by seed companies and protected by non-disclosure agreements, only raw estimations can be given for the size of the datasets. The total number of datasets present in the database is more than $1,000,000$. There are more than $50,000$ genomic, environmental and phenotypic variables (the n in 3.1), although only a limited number (10) of those variables were used as constraints in the following experiments.

4.2 Experiments

In the following figures, a sample of the most representative results are shown.

Figure 2 shows the convergence speed of the tolerance of a single constraint upon various experiments. It exhibits that a limited number of steps is needed to reach a fixed point, according to the constraints strength. The tolerance converges to different values due to the fact that this particular constraint may be of more or less importance depending on the problem. It can be seen that the tolerance evolves by stages. This pattern can be explained by the fact that the Constraint Agent tightens its constraint only if the number of Inactive+Active data still matching the new constraint is sufficient. As this number steadily increases over time, the constraint can be tightened only when a certain threshold is reached. For example, for Experiment 2, the tolerance remains constant from step 105, which means that from this step on, the Constraint Agent related to this constraint decides at each iteration to leave the tolerance as is. However, the other constraints –not shown in this figure– are still able to adjust their tolerance. To see when the fixed point is actually reached, the analysis presented in Fig.3 is necessary.

Figure 3 shows the total number of data used against the simulation time, in iteration steps. For example, for experiment 2, the fixed point is reached at 108 steps. Those results exhibit that less than 1% of the database is needed for the system to reach its fixed point and return a prediction to the user in less than 200 steps. An experiment runs in about 45 minutes, however we estimate that more than 75% of this time is consumed by database accesses. More precise measurements have still to be made.

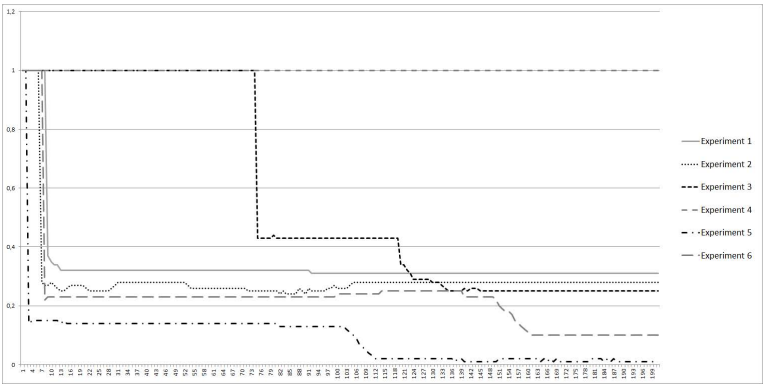

Fig. 2. Convergence of a single constraint upon various experiments

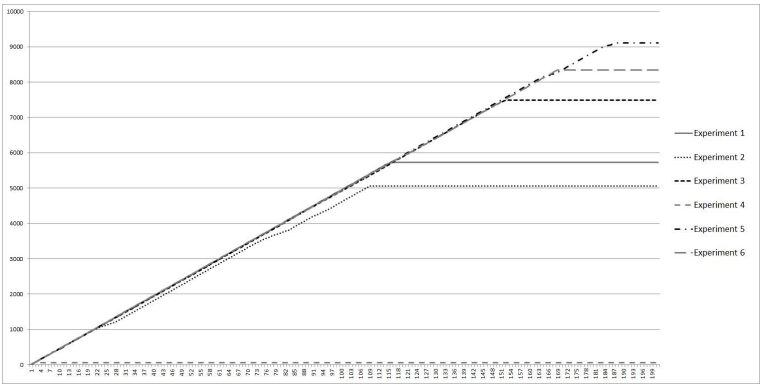

Fig. 3. Convergence of the Inactive+Active dataset size upon various experiments

5 Conclusion

In this paper, an Adaptive Multi-Agent System was presented to overcome the lack of traditional statistical approaches in phenotypic prediction. The system

solves the problem using cooperation between agents, which are responsible for the various genomic and environmental constraints. Experiments show how the system converges towards a solution despite the high sparsity of the data involved. Since only the relevant datapoints are explored based on a very small fraction of the entire database, the system is not sensitive to missing data. The system convergence is characterized by the convergence of constraints tolerance and the stability of the data pool.

Current and future works are aimed at validating the solution by different cross-validation tests as well as systematic comparison with current statistical methods.

Acknowledgements. This work is part of the GBds (Genomic Breeding decision support) project funded by the French FUI (Fonds Unique Interministeriel) and approved by Agri Sud-Ouest Innovation (competitive cluster for the agriculture and food industries in southwestern France). We would like also to thank our partners in this project (Ragt 2n, Euralis and Meteo France) and Daniel Ruiz from the IRIT laboratory for his help concerning data analysis.

References

[1] Food, of the United Nations, A.O.: State of food and agriculture (2013)
[2] Lorenz, A.J., Chao, S., Asoro, F.G., Heffner, E.L., Hayashi, T., Iwata, H., Smith, K.P., Sorrells, M.E., Jannink, J.L.: Genomic selection in plant breeding: Knowledge and prospects. Advances in Agronomy 110, 77–121 (2011)
[3] Moser, G., Tier, B., Crump, R., Khatkar, M., Raadsma, H.: A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide snp markers. Genetics Selection Evolution 41(1), 56 (2009)
[4] Whittaker, J.C., Thompson, R., Denham, M.C.: Marker-assisted selection using ridge regression. Genetical Research 75, 249–252 (2000)
[5] Hayes, B.J., Visscher, P.M., Goddard, M.E.: Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research 91, 47–60 (2009)
[6] Ferber, J.: Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence. Addison-Wesley Longman Publishing Co., Inc. (1999)
[7] Capera, D., George, J.P., Gleizes, M.P., Glize, P.: The AMAS Theory for Complex Problem Solving Based on Self-organizing Cooperative Agents. In: International Workshop on Theory And Practice of Open Computational Systems (TAPOCS@WETICE 2003), pp. 389–394. IEEE Computer Society (2003)
[8] Ilin, A., Raiko, T.: Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res. 11, 1957–2000 (2010)

NAPROC-13: A Carbon NMR Web Database for the Structural Elucidation of Natural Products and Food Phytochemicals*

José Luis López-Pérez[1], Roberto Theron[2], Esther del Olmo[1], Beatriz Santos-Buitrago[3], José Francisco Adserias[4], Carlos Estévez[4], Carlos García Cuadrado[4], David Eguiluz López[4], and Gustavo Santos-García[5]

[1] Departamento de Química Farmacéutica – IBSAL – CIETUS, Universidad de Salamanca, Spain
[2] Dpto. Informática y Automática, Universidad de Salamanca, Spain
[3] School of Computing, University of the West of Scotland, UK
[4] Fundación General Universidad de Salamanca, Spain
[5] Computing Center, Universidad de Salamanca, Spain

Abstract. This paper describes the characteristics and functionalities of the web-based database NAPROC-13 (`http://c13.usal.es`). It contains Carbon NMR spectral data from more than 21.000 Natural Products and related derivates. A considerable number of structures included in the database have been revised and corrected from the original publications considering subsequent published revisions. It provides tools that facilitate the structural identification of natural compounds even before their purification. This database allows for flexible searches by chemical structure, substructure of structures as well as spectral features, chemical shifts and multiplicities. Searches for names, formulas, molecular weights, family, type and group of compound according to the IUPAC classification are also implemented. It supports a wide range of searches, from simple text matching to complex boolean queries. These capabilities are used together with visual interactive tools, which enable the structural elucidation of known and unknown compounds by comparison of their ^{13}C NMR data.

Keywords: structural elucidation, carbon NMR spectral database, natural compounds, chemoinformatics, bioinformatics, food phytochemicals, SMILES code.

1 Introduction

Chemoinformatics is the application of informatics methods to chemical problems [7]. All major areas of chemistry can profit from the use of information technology and management, since both a deep chemical knowledge and the

* Financial support came from the Ministerio de Educación y Ciencia, project TIN2006-06313 and the Junta de Castilla y León, project SA221U13. Research also supported by Spanish project Strongsoft TIN2012-39391-C04-04.

J. Sáez-Rodríguez et al. (eds.), *8th International Conference on Practical Appl. of Comput. Biol. & Bioinform. (PACBB 2014)*, Advances in Intelligent Systems and Computing 294,
DOI: 10.1007/978-3-319-07581-5_2,

processing of a huge amount of information are needed. Natural Products (NPs) structure elucidation requires spectroscopic experiments. The results of these spectroscopic experiments need to be compared with those of the previously described compounds. This methodology provides highly interesting challenges for chemoinformatics practitioners.

NPs from microbial, plant, marine, or even mammalian sources have traditionally been a major drug source and continue to play a significant role in today's drug discovery environments [10]. In fact, in some therapeutic areas, for example, oncology, the majority of currently available drugs are derived from NPs. However, NPs have not always been as popular in drug discovery research as one might expect, since in the NPs research, tedious purifications are needed in order to isolate the constituents. These procedures are often performed with the main purpose of structure identification or elucidation. Because of that, ultra-high throughput screening and large-scale combinatorial synthetic methods have been the major methods employed in drug discovery [19]. Yet if the structures of natural extract constituents could be known in advance, the isolation efforts could be focused on truly novel and interesting components, avoiding re-isolation of known or trivial constituents and in this way increasing the productivity [4]. Furthermore, it is generally known that the intrinsic diversity of NPs exceeds the degree of molecular diversity that can be created by synthetic means, and the vast majority of biodiversity is yet to be explored [10]. At present, it is unanimously assumed that the size of a chemical library is not a key issue for successful developmental leads and that molecular diversity, biological functionality and "drug likeness" are decisive factors for drug discovery processes [10]. For this reason, the natural products-based drug discovery is on the rise again.

Some chemoinformatics methods include predictive classification, regression and clustering of molecules and their properties. In order to develop these statistical and machine learning methods the need for large and well-annotated datasets has been already pointed out. These datasets need to be organized in rapidly searchable databases to facilitate the development of computational methods that rapidly extract or predict useful information for each molecule [3]. The progressive improvement of analytical techniques for structural elucidation makes today's structural identification more reliable and it permits the correction of structures of a large number of previously published compounds.

NPs databases are of high priority and importance for structure search, matching and identification [9]. In this paper, we present a web-based spectral database that facilitates the structural identification of the natural compounds previous to their purification.

2 ^{13}C NMR Spectroscopy: A Power Technique for Structural Elucidation of Natural Compounds

For the elucidation of natural compounds, ^{13}C NMR spectroscopy is the most powerful tool. This is largely due to the well-known and exquisite dependence of the ^{13}C chemical shift of each carbon atom on its local chemical environment

and its number of attached protons. Furthermore, the highly resolved spectra, provided by a large chemical shift range and narrow peak width, could by easily converted into a highly reduced numerical lists of chemical shift positions with minimal loss of information. ^{13}C NMR spectroscopy can also provide the molecular formula. The analysis of spectral data for the determination of unknown compound structure remains a usual but a laborious task in chemical practice.

Since the advent of computers many efforts have been directed toward facilitating the solution to this problem [7]. Libraries of such spectral lists of data are common for synthetic organic compounds and are an invaluable tool for confirming the identity of known compounds [17]. However, the methods for structure elucidation of compounds apart from a database have not been exhaustively studied. In the field of NPs, where hundreds of thousands compounds have been reported in the literature, most compounds are absent from commercially available spectral libraries.

Once a researcher in NPs isolates and purifies a compound, he needs to know the compound's structure, the skeleton and if it has been previously described. If a database of NPs and their NMR spectral data are available, searching databases will allow for quick identifications by means of comparison of the new compound with the NMR spectrum of the registered compounds or with other related compounds. This search provides insight into the structural elucidation of unknown compounds.

NAPROC-13 has many search facilities and a set-up that allows comparative studies of related compounds. At present, new search tools are being developed and the data input methods are being improved so as to allow researchers from different institutions to introduce the information over the Net. The aim of this database is to help identify and elucidate the structure of hypothetical new compounds, by comparing their ^{13}C NMR data with those of already published related compounds.

2.1 NMR Databases for Phytochemicals

Mass spectrometry and NMR spectroscopy allow the efficient identification of phytochemicals and of other NPs. Because of the large spectral dispersion, the relative chemical shift invariance, and the simplicity of ^{13}C NMR spectra, most analytical chemists prefer to use ^{13}C NMR for the identification of phytochemicals, phytochemical metabolites, and other NPs. NAPROC-13, which is a ^{13}C NMR database of NPs, probably represents one of the richest NMR resource for phytochemists and phytochemical databases [18]. Along with NAPROC-13, NMRShiftDB2 [20] (`http://www.nmrshiftdb.org`) is another open web database for organic structures and their NMR spectra; unfortunately, it does not contain too many NPs. Other noteworthy NMR databases are: HMDB, HMDB, MMCD, BMRB, SDBS, and HaveItAll CNMR-HNMR.

The Human Metabolome Database [25] (HMDB, `http://www.hmdb.ca`) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. This database contains >40.000 metabolite entries. It contains experimental ^{1}H and ^{13}C NMR data (and

assignments) for 790 compounds. Additionally, predicted ^{1}H and ^{13}C NMR spectra have been generated for 3.100 compounds.

Spectral Database for Organic Compounds (SDBS, `http://sdbs.db.aist.go.jp`) is an integrated spectral database system for 34.000 organic compounds, which includes 6 different types of spectra (an electron impact Mass spectrum EI-MS, a Fourier transform infrared spectrum FT-IR, a ^{1}H NMR spectrum, a ^{13}C NMR spectrum, a laser Raman spectrum, and an electron spin resonance ESR spectrum) under a directory of the compounds.

HaveItAll CNMR-HNMR Library (`http://www.bio-rad.com`) access over 500.000 high-quality ^{13}C NMR and 75.000 ^{1}H NMR spectra. It offers access to high-quality NMR spectral reference data for reliable identification and NMR prediction.

3 NAPROC-13: Database and Web Application

The structural elucidation of natural compounds poses a great challenge because of its great structural diversity and complexity [16]. For this reason, we are developing a database accessible through a standard browser (`http://c13.usal.es`). It provides the retrieval of natural compounds structures with ^{13}C NMR spectral data related to the query. At present it contains the structures of more than 21.000 compounds with their ^{13}C NMR information.

MySQL (`http://www.mysql.com`) has been chosen to develop NAPROC-13 for its high reliability and good performance; it is a fast, robust multithread, multiuser database. MySQL is an open-source relational database manager, based on SQL (Structured Query Language). We use the open-source Apache Tomcat web server and JavaServer Pages (JSP) technology to create dynamically web pages. By means of proper JSP programming, we bring about the communication between applets and the database. As for the interactive visualization tools, Java applets have also been integrated in this application.

There is a widespread belief that publicly funded scientific data must be freely available to the public [18]. Open accessibility has many benefits, not the least of which is increased visibility. Our database makes freely available resources that can be easily accessed over the Internet without passwords or logins.

Our aim was to design a reliable database. Data acquiring are fully and properly provided with references, data sources, and citations. References ensure that the data can be reproduced and allow users to investigate the data sources for further information. Structures and spectral data collected in the database proceeds from books, journals and our measurements. They are mainly compiled from papers in the following research journals: *Journal of Natural Products, Phytochemistry, Planta Medica, Chemical & Pharmaceutical Bulletin, Chemistry of Natural Compounds, Helvetica Chimica Acta*, and *Magnetic Resonance in Chemistry.*

NAPROC-13 is continually expanded and updated in order to enhance the database's querying capabilities, design, and layout. User-friendliness has been another important factor. NAPROC-13 interface allows for complex queries,

which can be performed through simple pull-down menus or clickable boxes using plain language. Web capabilities of HTML language enables a high degree of interactivity.

4 Reliability of Structural and Spectroscopic Data from NAPROC-13

Over the course of the past four decades, the structural elucidation of unknown NPs has undergone a tremendous revolution. A battery of advanced spectroscopic methods, such as multidimensional NMR spectroscopy, high-resolution mass spectrometry or X-ray crystallography are available today for structural resolution of highly complex natural compounds.

Despite the available structural strategies and elucidation methods and despite the progress made in recent decades, constant revisions of structures of NPs are published in the literature. These revisions are not mere inversion of stereocenters, but they may entail profound changes in their chemical constitution. For example, more than thousands of articles on structural revisions published in the best journals cover virtually all types of compounds, steroids, terpenoids, alkaloids, aromatic systems, etc., regardless of the size of the molecule.

Often a structure obtained by X-ray diffraction is considered irrefutable proof of its structure. However, we can find examples in which the position of 2 heteroatoms has been changed. Another method to confirm the structure of a compound is by means of its total synthesis. In some of the synthesized compounds, we can observe a discrepancy between the natural product and its synthetic data, which means that the proposed structure for the natural compound is not correct. Although in most cases, the structure can be fixed, in others, the ambiguity persists since the NMR data of the synthesized compound is different from the structural proposal of the natural product and, hence, the actual structure remains unknown. This is due to the enormous structural complexity of the isolated compounds and the small quantities of sample available.

In the field of NPs, the structural assignment is often based on the structures of the related compounds. Thus, if the wrong structure of a substance is taken as a model, errors are continuously replicated. This problem can be avoided, if reliable NPs spectroscopic data is entered into a database such as NAPROC-13 and is used as reference. In this way, we can avoid some errors in publications. Let's consider the following example: the same compound was independently isolated by two research teams who propose different structures and names for the same spectroscopic data of an identical compound. Later, both structural proposals are proven to be incorrect (see Figure 1) Access to the spectroscopic data could help in assigning new compounds of a family and facilitate the process of structural reallocation.

Incorrect NPs assignments not only make the determination of the biosynthetic pathway more difficult, but may have costs in terms of time and money. Imagine that an interesting product is isolated from a pharmacological point of view. Current strategy synthesizes NPs and their closely related analogues.

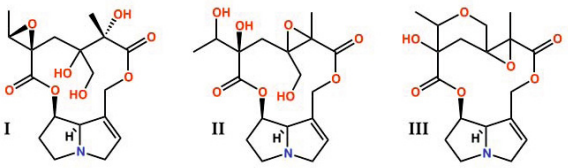

Fig. 1. I: Proposed erroneous structure in [23]; II: Proposed erroneous structure in [1]; III: Corrected structure of adenifoline [26]

Obviously, if the structure is not correct, we synthesize another compound different from the one we are interested in.

A database is as useful as the data it contains. Curators spend a considerable amount of time acquiring data in order to keep the database relevant. Data acquisition and data entry are not automated, but data is manually searched, read, assessed, entered, and validated. NAPROC-13 prioritizes the introduction of those compounds whose structures have been reviewed in recent literature. Since a database of this nature grows, manual transcription errors and those present in the literature are inevitable, We have developed some scripts to detect obvious chemical shift errors, such as shifts greater than 240.0 ppm, as well as errors based on a few simple rules regarding proper ranges of chemical shift ranges for several easily identifiable functional groups. Thus the data presented in NAPROC-13 has greater reliability when being considered as a pattern.

4.1 Database Design

Numbering system of the well-known *Dictionary of Natural Products* (`http://dnp.chemnetbase.com`, Chapmann & Hall/CRC Press) has been applied to each family skeleton in NAPROC-13. Numbering homogeneity within the same family compounds enables the comparison of spectral data for a variety of related structures.

NAPROC-13 contains a wide diversity of data types. It collects a rich mixture of text, numbers, charts, and graphs. The basic database schema is relationally organized and the molecular structures are defined and stored in the database with SMILES code (Simplified Molecular Input Line Entry Specification) [24]. This format of structural specification, that uses one line notation, is designed to share chemical structure information over the Internet [8]. For example, SMILES code for Melatonin ($C_{13}H_{16}N_2O_2$) is "`CC(=O)NCCC1=CNc2c1cc(OC)cc2`".

Substructural searches are performed by SMARTS specification (SMiles ARbitrary Target Specification). This is a language for specifying substructural patterns in molecules from the SMILES code. The SMARTS line notation is expressive and allows extremely precise and transparent substructural specification and atom typing. It uses logical operators that allow choosing all-purpose atoms, groups of alternative atoms, donor and acceptor groups of hydrogen bonds or lipophilic atoms. For example, SMARTS specification for Hydrazine (H_2NNH_2) is "`[NX3][NX3]`" and for an Oxygen in $-O-C=N-$ is "`[$([OX2]C=N)]`".

Evidently the type of patterns and notations that are used, both SMARTS and SMILES, are too complex to be interpreted by organic chemists without specific training in this area. For this reason, we use a tool able to convert these notations into a graph that represents a substructure that will act as question.

The spectral ^{13}C NMR data, in form of a numerical list of chemical shift and their multiplicity, is always associated with each compound structure. A script calculates and represents the ^{13}C NMR spectra of the selected compound in a very similar way to the experimentally obtained data, and shows the decoupled proton (broad band) and the DEPTs (Distortionless Enhancement by Polarization Transfer). Figure 2 displays the ^{13}C NMR spectrum calculated for the substance of a compound found by a search. Another script calculates and represents the signals corresponding to the deuterated solvent used in the experiment.

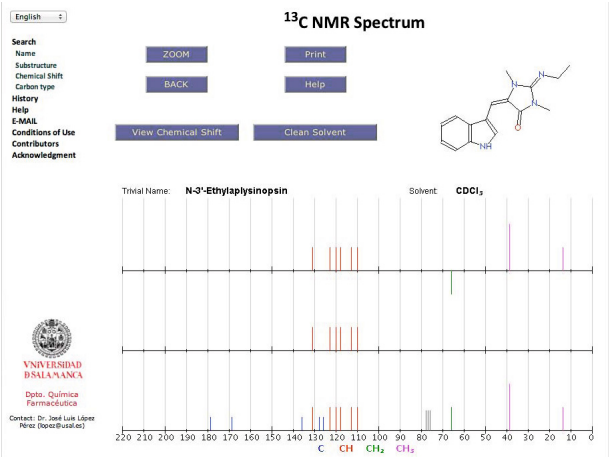

Fig. 2. Chart of the ^{13}C NMR spectra of a chemical compound. Multiplicities of the carbons are codified by colors.

5 Queries in NAPROC-13

NAPROC-13 allows for flexible searches by chemical structure, substructure of structures as well as spectral features, chemical shifts and multiplicities [21]. Searches for names, formulas, molecular weights, family, type and group of compound according to the IUPAC classification and other parameters are also included. NAPROC-13 database supports a wide range of searches, from simple text matching to complex boolean queries.

This database offers several alternatives of the chemical shift search process. The multiplicity for each chemical shift is always required and this constitutes a useful search restriction. The search can be undertaken for one specific position in the molecule. The system permits to formulate the enquiry with the required number of carbons, by one carbon or more, up to the totality of the carbons

of the compound. There is a default established deviation (+/-1 ppm) for all chemical shifts, but the user can specify a particular deviation for every carbon. It is important to be able to repeat the search with different deviations and to select the search that provides the best results. If the deviation is too small, it may occur that an interesting compound will not be selected. In this way, a reasonable and manageable number of compounds can be obtained. Even a search based only on the most significant carbons of the studied compound ^{13}C NMR spectrum will lead to the identification of the family they belong to.

Moreover, users can address questions in a graphic form to the database using JME Molecular Editor, a structure editor that enables the user to draw several fragments that may not be related to each other. JME has a palette that speeds up the creation of structures and uses IUPAC recommendations to depict the stereochemistry. By using this palette it is possible to add preformed substructures, i.e., different size cycles, aromatic rings, simple and multiple bonds, frequently used atoms. The control panel allows to enter directly functional groups, i.e., carboxyl acids, nitro groups and other groups. The facilities of this applet rapidly generates a new structure and speeds up the search process.

It is also possible to undertake a combined and simultaneous search by substructure and by chemical shifts, a feature that undoubtedly enhances the search capacity and increases the possibilities of finding compounds related with the problem substance.

The iterative search is probably the most genuine search of this application. The user can include in his search from one chemical shift to the totality of the signals of the ^{13}C NMR spectrum problem compound. This tool will initially carry out a search of all the entered chemical shifts. If it does not find any compound that does not fulfill the full requirements, it will undertake a new iterative search by all the shifts except one. It will perform all the possible combinations until it finds a compound that fulfills some of the requirements.

The matching records retrieved resulting from a search can be displayed in the Results pane in the form of molecular structure. The chemical shifts of the matching records can be viewed in tables or in the compound structures by clicking the δ (ppm) in tables/structures buttons. Properties pane provides the details of a particular record. Spectrum pane shows spectrum graphically.

5.1 Interactive Visual Analytical Tool

As stated above, NAPROC-13 features a built-in visual analytical tool. It is a highly interactive interface integrated by four linked views: ^{13}C NMR spectrum, structure, parallel coordinates plot, and taxonomic information (see Figure 3).

The main advantage of this approach is that the user can deal with a great number of compounds that have matched a particular search. Thanks to interaction, a user can explore this result set, focusing on particular details of a given compound (name-family-type-group, structure, spectrum) while maintaining the context, i.e. the characteristics of the rest of the compounds in the result set. Parallel coordinates provide a way of representing any number of dimensions in the 2D screen space [11]. Each compound is drawn as a polyline passing through