Protein Engineering, Design & Selection vol. 23 no. 12 pp. 899–902, 2010
Published online October 1, 2010 doi:10.1093/protein/gzq068

SHORT COMMUNICATION

CDMS (Clinical Data Mining Software): a cytokinome data mining system for a predictive medicine of chronic inflammatory diseases

D.Evangelista[1,2], G.Colonna[3], M.Miele[1,3], F.Cutugno[2], G.Castello[1], S.Desantis[4] and S.Costantini[1,3,5]

[1]Centro Ricerche Oncologiche Mercogliano, 'Fiorentino Lo Vuolo', via Ammiraglio Bianco, 83013 Mercogliano, Avellino, Italy, [2]Dipartimento di Scienze Fisiche, Università Degli Studi di Napoli Federico II, Campus di Monte S. Angelo, Via Cintia 25, 80125 Napoli, Italy, [3]Dipartimento di Biochimica e Biofisica & Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università di Napoli, via Costantinopoli 16, 80138 Naples, Italy and [4]Altravia srl, Via Andrea Millevoi 683, 00178 Roma, Italy

[5]To whom correspondence should be addressed.
E-mail: susan.costantini@unina2.it

Received July 5, 2010; revised August 31, 2010; accepted September 7, 2010

Edited by Leo James

The cytokines, main players of the chronic inflammation progression leading to serious diseases such as diabetes or cancer, represent a target for better clinical prognosis and innovative therapeutic strategies. To investigate the immunopathogenetic progression of these diseases, the evaluation of serum cytokines profiles made of many different proteins is much more informative than single protein measurements. We developed a Clinical Data Mining Software to collect cytokine profiles evaluated on healthy subjects and patients by multiplex immunoassays also annotated with their clinical and laboratory data, to compare patient profiles by statistical tools and to evaluate their disease progression.
Keywords: cytokine profile/data mining

Introduction

The progressive increase in electronically stored clinical data is opening the possibility of carrying out large-scale studies aimed to discover correlations between new research data and related diseases. For these reasons, many relational databases have implemented data mining techniques (Harrison, 2008). Data mining has been described as the 'extraction of implicit, previously unknown and potentially useful information', such as associations and correlations between data elements from large repositories of data (Lee and Siau, 2001). Data mining techniques include methods to group data elements that have similar features (association rules, clustering etc.) and methods to determine predictive relationships between groups of data elements (classification and neural network) (Hand, 2007). An overview of open-source data mining tools is provided by Zupan and Demsar (Zupan and Demsar, 2008) explaining that there are diverse data mining methods and user interfaces. Moreover, they demonstrate that this field and its tools are ready to be fully exploited in biomedical research.

However, the scientific community needs clinical laboratory databases to collect medical data related to diseases progression and therapy response. In the last years, particular attention has been focused on the protein class comprising cytokines, chemokines and growth factors, because they play a crucial role in promoting angiogenesis, metastasis and subversion of adaptive immunity. On the whole, we define with the term 'cytokinome' the totality of these proteins and their interactions in and around biological cells. In particular, these proteins are involved in cancer-related chronic inflammation and can represent a target for innovative clinical prognosis useful for therapeutic and clinical strategies (Mantovani and Pierotti, 2008; Mantovani *et al*., 2008). Since the control of cytokine production is highly complex and multifactorial, their effects are mediated through multiple regulatory networks. The intricate complexity of these networks clearly conceals the role that a single cytokine may play in the pathogenesis of the disease. Therefore, it is more informative to investigate the immunopathogenesis of a disease process by analyzing a multiple panel of cytokines (Costantini *et al*., 2009). Utilizing a bead-based broad-spectrum multiplex immunoassay, it is possible not only to evaluate the serum levels of those cytokines ensemble that effectively correlate with the progression of the disease activity but also to define the immunomodulatory effects of a therapy even after months of treatment (Sato *et al*., 2009; Ozturk *et al*., 2009; Capone *et al*., 2010). This indicates that the definition and evaluation of a human cytokinome represents an important future tool to analyze the interaction network of cytokines both in healthy individuals and in patients affected by different diseases. In fact, it will permit one to understand and investigate how the regression of a chronic inflammation process, by acting on the cellular populations of cytokines, can block the progression of a cancer and, therefore, it can be a useful prognostic and diagnostic tool for clinicians.

For these reasons, a portal with user-friendly interfaces, which can be used both by physicians and researchers not only to collect and to correlate clinical data and serum levels of cytokines but also to know quickly what cytokines, chemokines or growth factors are significant in the progression state of a given disease, represents an important and useful tool for clinical prognosis and therapy studies.

In this work, we have developed software to collect clinical data and serum levels of many cytokines, chemokines and growth factors evaluated on healthy subjects and patients

affected by different diseases using multiplex immunoassays. Moreover, some statistical tools were implemented to correlate significatively clinical and experimental data. At present, we focused our attention on chronic hepatitis C and hepatocarcinoma (HCC) diseases. However, our software can be easily upgraded for other chronic diseases. We are also developing a routine to quickly compare standardized cytokinome profile of a patient against a whole data bank that collects cytokinome data from some different diseases. This routine will be able to support a reliable diagnosis/prognosis of patients with a progressive disease clinically still silent. Of course, the more data there are in the data bank, the more reliable will be the result. We think that this type of tool is essential for a predictive medicine. Our portal is named CDMS (Clinical Data Mining Software) and is accessible at the URL: http://www.cro-m.eu/CDMS/.

Methods

CDMS was developed using Web server Apache/2.2.11, MySQL version 5.0.75, PHP/5.2.6-3 Ubuntu version 9.04 and phpMyAdmin version 3.1.2 deb1 Ubuntu 0.2. It is a web-oriented software that was realized using the scripting language PHP, the JavaScript technology for dynamic contents, the markup language HTML and style sheet CSS 2.0. For most of the statistical analysis, the aggregate functions of MySQL were used. Moreover, we developed specific PHP scripts for the *t*-test formula, correlation formula and related graphs. Block diagram of portal functionality is shown in Fig. 1.

Description of CDMS portal

CDMS software is used for the collection of clinical data and cytokine concentrations evaluated by broad-spectrum bead-based multiplex immunoassay in healthy donors and in patients with chronic hepatitis C and HCC. In this portal, there are some statistical tools to analyse both clinical data and serum levels of cytokines and to visualize these data through graphs. CDMS was developed using the web-oriented typology that permits (i) the consultation, updating and processing of all information related to each patient in real time and in remote modality, (ii) the information sharing between authorized users and (iii) data security.

CDMS allows certified users to access some of its services on the basis of their privileges. In detail, homepage for the administrator presents *users administration*, *patient administration* and *statistical analysis* sections. Physicians and Researchers can access the patient administration and statistical analysis sections, and all other authorized figures can access only statistical analysis section. In the patient administration section, there are case histories of patients with information related to their diagnosis, biological analyses as well as clinical data, and cytokine evaluations. In particular, in this section, one can insert 122 biological data for each patient (Table I) and the concentrations of 50 cytokines comprising 20 interleukins, 12 chemokines, 2 interferons, 10 growth factors and other 6 proteins (Table II). In detail, these cytokines were chosen because they are present in two commercial kits (BioRad Lab., Inc.) used to evaluate the cytokines serum levels in blood samples from healthy subjects and patients with chronic hepatitis C or HCC by using multiplex immunoassays. In fact, many of these molecules (i.e. IL-8, CXCL9, CXCL10) are the most representative in these two diseases. However, CDMS can be easily upgraded and, therefore, it is possible to insert both other chronic diseases and other proteins. Moreover, for the same patients, it is possible to insert the cytokine profiles evaluated at different times to compare and evaluate results at different stages of the disease.

In the statistical analysis section, the user can select the disease, filter the patients on the basis of gender, age and experiment date and select the most appropriate tool to perform the statistical analysis. In particular, we have implemented: (i) median, mean, variance, standard deviation, min and max values for the selected protein (Supplementary data, Fig. 1SA); (ii) *t*-test value related to the comparison between cytokine concentrations in control group and patients; (iii) Pearson correlation between different cytokines with related graph (Supplementary data, Fig. 1SB); (iv) Pearson correlation between each cytokine and some clinical data (i.e. tumor size) with related graph.

The system administrator can access to the users administration section for all the registration procedures of new users assessing 'username and password' and mailing them to users. The password will be created in random modality by using the MD5 algorithm (Rivest, 1991) and modifiable at user's request. All the authorized users can access the glossary section for information and details related to biological and clinical meaning of cytokines.

Conclusions

The relationships between chronic inflammation and progression of diseases represent a topic of great interest but one that is not yet used for a reliable predictive medicine because of its inherent complexity due to the very large amount of data to take into account. In fact, it is well known that the principal players of the chronic inflammatory processes are the cytokines, chemokines and growth factors but it is still difficult to analyze the whole spectrum of their progression in time without an adequate informative support. Therefore, CDMS represents, to the best of our knowledge, the first 'user-friendly' tool that can be used by researchers as well as physicians and clinicians to significatively correlate clinical data and cytokine profiles and to identify what cytokines can be significant for the examined disease at a given time. In particular, we have used CDMS in our recent work (Capone *et al*., 2010) where the HCC was chosen as a studying model. In fact, the serum levels of 50 different cytokines, chemokines and growth factors were evaluated in patients affected by HCC with chronic HCV-related hepatitis and liver cirrhosis using a bead-based broad-spectrum multiplex immunoassay. Both clinical data and cytokines serum levels of all the patients and of healthy subjects group, included as controls, are collected in CDMS. Using the available statistical tools, we have identified the cyto-chemokines pattern involved in the chronic inflammation processes versus HCC (i.e. IL-1α, IL-6, IL-8, IL-12p40, GM-CSF, CCL27, CXCL1, CXCL9, CXCL10, CXCL12, β-NGF). Moreover, it has been possible by CDMS to correlate the cytokine serum levels with the clinical data and to verify that IL-8 correlates significantly with large tumor size (>5 cm), and it can be used both as a useful marker of tumor invasiveness and as an

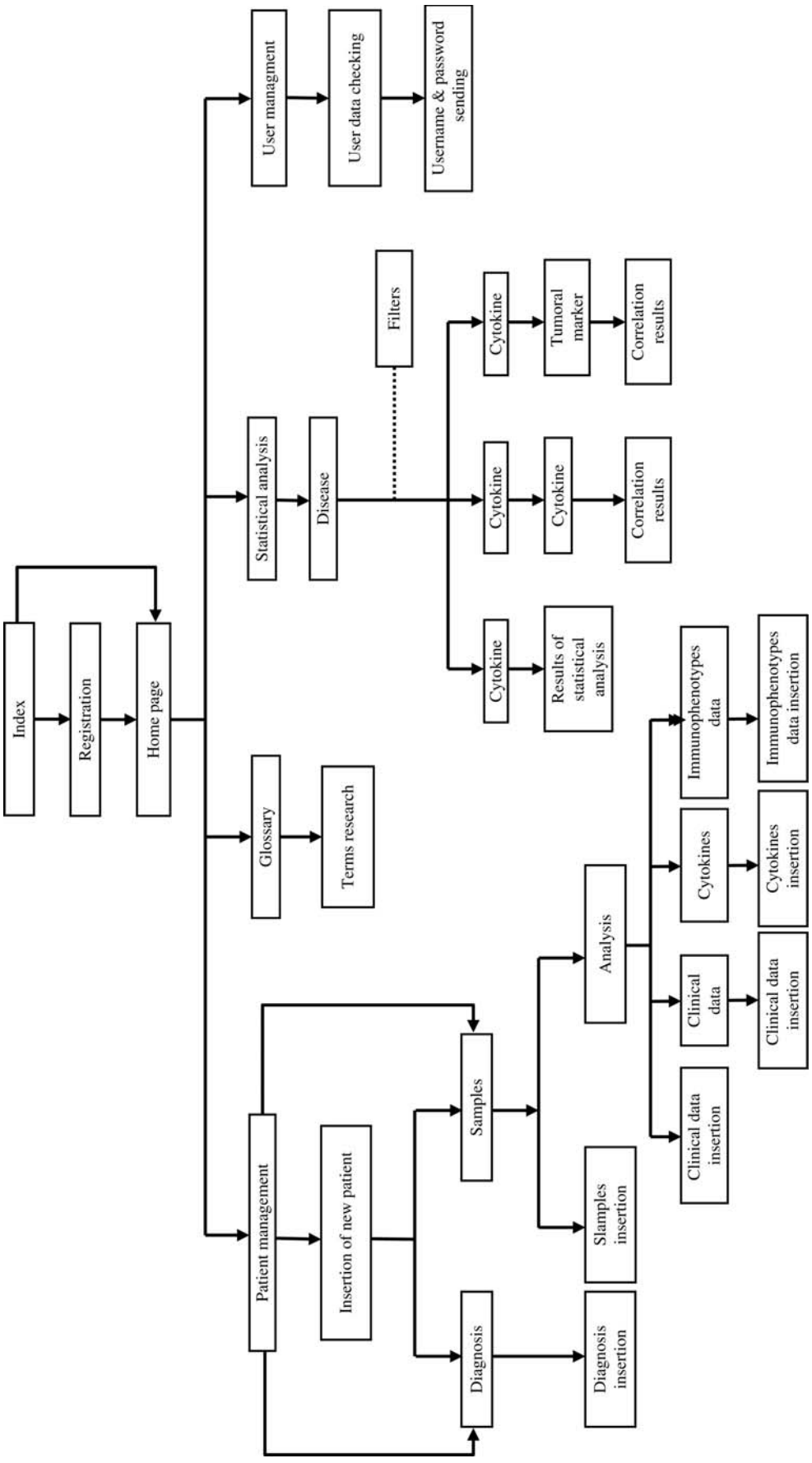

Fig. 1. Block diagram of the software functionality

Table I. List of data for patient administration section

Personal data and diagnosis			
Inserting date	Id diagnosis	Diagnosis	Etiology
Name	Id patient	Therapy type	State (live/dead)
Surname	Diagnosis date	Chronic hepatitis	Motivation dead
Gender	Operation date	Cirrhosis	Clinical presentation
Place of birth	Last control date	Child	Type
Date of birth	Follow-up	No cancers	Side
Blood	Previous treatment date	Cancer place	Tumor size
Rh-Positive	Previous treatment	Cancer dimension	Lymph nodal status
Building	Treatment	Metastasis	Cellular type
Healthy	Treatment date	TNM	Sarcomatoid variant
Notes	Surgical technique	Grading	Diagnosis note
Clinical data			
AbT	Got	CEA	MON
AbTPO	GPT	CIC	Na
Ac. Urico	HAV	Cl	NEU
αFP	HBV	Col.Tot	PLT
Albumin	HCV	Colinesterase	PM
ALP	HGB	Cortisol	PRL
ALT	IgA	Free cortisol	Proteins
Amilase	IgG	CPK	PT
Antibodies to adrenal cortex	Igk	Creatinine	TE
Antibodies to pituitary	Igl	EOS	PTT
AST	IgM	Estradiol	RBC
Azotemy	IL-6	Fe mc	sIL2-R
β2M	INR	FOS. ALCALINE	One
BAS	k/l	FSH	HCV title
Bilirubin Tot.	k/l	fT3p	TNFα
Bilirubin Dir	LDH	fT4p	Tryglicerides
C3	LH	Genotype HCV	BOSIS
C4	LUC	gGT	TSH
CA	LYN	Glycemia	WBC
CA19-9	Mg		

Table II. List of Cytokines

Interleukins	IL-1α, IL-1β, IL-1ra, IL-2, IL-2R, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-12p40, IL-12p70, IL-13, IL-15, IL-16, IL-17, IL-18
Chemokines	MCP-1, MCP-3, MIF, MIP-1α, MIP-1β, RANTES, Eoxatin, CTACK, CXCL1, CXCL9, CXCL10, CXCL12
Growth factors	Basis FGF, β-NGF, G-CSF, GM-CSF, HGF, M-CSF, PDGF-ββ, SCF, SCGF-β, VEGF
Interferons	INF-α2, INF-γ
Other proteins	LIF, ICAM-1, TNF-α, TNF-β, TRAIL, VCAM-1

independent prognostic factor for HCC patients (Capone *et al*., 2010).

Therefore, this tool can be a useful support to develop a reliable predictive medicine. Moreover, we are planning to open the data set to other diseases and implement other statistical tools and classification methods to improve or discover new predictive relationships among data groups.

Supplementary data

Supplementary data are available at *PEDS* online.

References

Capone,F., Costantini,S., Guerriero,E., Calemma,R., Napolitano,M., Scala,S., Izzo,F. and Castello,G. (2010) *Eur. Cytokine Netw.*, **21**, 99–104.

Costantini,S., Capone,F., Guerriero,E. and Castello,G. (2009) *Immunol. Lett.* **126**, 91–92.

Hand,D.J. (2007) *Drug Saf.*, **30**, 621–622.

Harrison,J.H. (2008) *Clin. Lab. Med.*, **28**, 1–7.

Lee,S. and Siau,K. (2001) *Ind. Manage. Data Syst.*, **100**, 41–46.

Mantovani,A. and Pierotti,M.A. (2008) *Cancer Lett.*, **267**, 180–181.

Mantovani,A., Romero,P., Palucka,A.K. and Marincola,F.M. (2008) *Lancet*, **371**, 771–783.

Ozturk,B.T., Bozkurt,B., Kerimoglu,H., Okka,M., Kamis,U. and Gunduz,K. (2009) *Mol. Vis.*, **15**, 1906–1914.

Rivest,R. (1991) in Menezes,A.J. and Vanstone,S.A. (eds), *Advances in Cryptology—CRYPTO '90 Proceedings*. Springer-Verlag, 303–311.

Sato,T., Kusaka,S., Shimojo,H. and Fujikado,T. (2009) *Ophthalmology*, **116**, 2165–2169.

Zupan,B. and Demsar,J. (2008) *Clin. Lab. Med.*, **28**, 37–54.