Università degli Studi di Torino

Facoltà di Scienze Matematiche Fisiche e Naturali

Dipartimento di Scienze Oncologiche

Dottorato di Ricerca in
Sistemi Complessi in Biologia Postgenomica

CICLO XX

FUNCTIONAL CLASSIFICATION OF ESTROGEN-RESPONSIVE GENE REGULATORY SEQUENCES IN BREAST CANCER CELLS: TOWARDS THE IDENTIFICATION OF REGULATORY NETWORKS

tesi presentata da:

Dott.ssa Gioia Altobelli

Relatori interni: *Proff. Michele De Bortoli - Michele Caselle*

Relatori esterni: *Dr. A. Benecke, Dr. G. Pavesi, e Prof. S. Subramaniam*

Coordinatore del ciclo: *Prof. F. Bussolino*

Anni Accademici: 2004 - 2007

SETTORE SCIENTIFICO-DISCIPLINARE DI APPARTENENZA:

BIO-11

Summary

Understanding regulation of estrogen-responsive genes is central to molecular biology and of great interest in medicine. The transcriptional activation/repression due to estrogen stimuli is gene- and cell- type specific, with a relevant molecular syntax being either unknown or not completely understood. Tissue- and cell type-specificity of the physiological response to estrogen has been addressed in experimental models by employing large-scale approaches, and results suggest that both complexity of transcriptional coregulators and epigenetics of chromatin organization are involved. Existence of several regulatory classes, e.g. early/late up/down- regulated clearly appears in microarrays and ChIP-on-Chip studies, but little is known about the underlying features of the corresponding gene regulatory sequences. Biochemical pathways target different DNA sequence elements and build up the combinatorial control which is a key in the regulatory events. The distribution of these DNA elements in the responsive gene regulatory regions should enable the inference of this regulatory networks.

We collected and compared expression data from genome-wide experiments in breast cancer cell models with a view to characterizing DNA flanking regions of hundreds of estrogen-responsive genes, possibly assessing differences between up- vs. down-regulated classes. In other words, we aimed at identifying the sequence motifs that may describe the differences between genes that are up- and down- regulated by estrogen, suggesting context features and possible control pathways. We set up a bioinformatics pipeline which combines traditional approaches focused on DNA sequence analysis of prox-

imal regions with a method that enables investigation of distal conserved nucleotide blocks of co-regulated estrogen-responsive genes (early responders only). We mainly focused our attention on those motifs identified by all of the tools and/or in different experimental datasets, with a view to inferring both regulatory factors to be tested in laboratory and relevant regulatory networks.

Although chromatin is a major context determinant of gene responsiveness, our pipeline handles the DNA sequences as linear strings. A topographic perspective is achievable to some extent, but we did not attempt it on a large scale. This pipeline also assumes that transcription binding sites tend to be overrepresented and to cluster in modules, and that evolution conservation is a key in their functionality. Despite universality of these assumptions has been recently challenged, we could assess remarkable features of the sequences which may have important biological implications. The upstream regions of early up-regulated genes strongly differ from the ones of early down-regulated genes, suggesting different regulatory mechanisms for the two classes of genes. Significant motifs' localization is provided, along with ontological analysis of gene subsets and transcription factor distributions. An example of co-localization of transcription factor binding sites in the 5'-flanking sequence of cyclin G2, which suggests a direct interaction between estrogen receptor and GATA-3 factor (ER-GATA), is also discussed in detail. This interaction may be important in mammary gland development.

Univeristà degli Studi di Torino

PhD course in Complexity in post-genomic Biology

XX cycle

A novel algorithm for supervised learning in neuronal models with binary synapses

Carlo Baldassi

Advisors:
prof. Nicolas Brunel
prof. Riccardo Zecchina

Introduction

The ability to dynamically adapt to the external stimuli and to retain the memory of past events are among the brain's most striking and crucial features, and are one of the most active fields of past and current research, both on the theoretical and on the experimantal side. Both processes of learning from experience and of memory formation are widely believed to occur through mechanisms of synaptic plasticity, i.e. of modulations of the signal transmission between neurons. However, due to the huge degree of complexity of the processes involved, describing their properties is a major challenge for both theroeticians and experimentalists.

In fact, an established framework about the synapses' dynamics is still lacking, despite the huge amount of experimental data collected, and many aspects of brain computations are yet unclear, including the signal encoding and whether the synaptic efficacies have a discrete or continuous nature.

On the other hand, at least some of the modifications induced through synaptic plasticity have to be permanent, while the biological environment in which these processes happen is subject to a very high level of noise; thus, the existence of a discrete set of stable states in a synapse would significantly improve its robustness. Multistability could be induced by positive feedback loops in protein interaction networks of the post-synaptic density, the small and highly specialized structure which is found in the dendritic spines [**18, 35, 6**]. This is in agreement with some recent experiments, which have suggested single synapses could be similar to noisy binary switches [**26, 23**], meaning that each synapse would have only two states, one with high conductance and one with low conductance.

From the theoretical point of view, however, there is an important difference between models which use continuous synaptic efficacies and those which use binary synapses, since it is in general much easier to develop efficient and plausible learning protocols in the continuous case, both in the unsupervised learning scenario (in which synaptic modifications are only induced by the pre and post-synaptic activities) and in the supervised scenario (in which an external 'teaching' or 'error' signal is present).

In fact, it has been shown [**32, 4, 5, 11**] that the performance of binary synapses systems (in terms of information stored per synapse) in the unsupervised scenario is very poor, unless two conditions are met: (1) activity in the network is sparse (very low fraction of neurons active at a given time); and (2) transitions are stochastic, with in average a balance between up and down transitions. This poor performance has motivated further studies [**12**] in which hidden states are added to the synapse in order to provide it with a multiplicity of time scales, allowing for both fast learning and slow forgetting. The hidden synaptic states are not directly involved in the unit's electrical properties, but rather they influence its plasticity properties; modifications of the hidden states are thus called "meta-plastic". As for the visible synaptic states, the hidden states could be represented by stable points of a protein interaction network.

In the supervised learning scenario, for the prototypical network in which this type of learning has been studied, the one-layer perceptron which has to perform a set of input-output associations, no efficient algorithms are known to exist when synapses have a finite number of states, in the case the number of input-output associations to be learned scales with the number of synapses. In fact, while learning in systems with analog synapse can always be achieved with simple algorithms if a solution exists, learning in systems with binary synapses is known to be a NP-complete task [**2, 3**], meaning that in the general case it belongs to the hardest computational class of combinatorial optimization problems. Moreover, even the easier case in which the patterns which have to be classified are supposed to be generated at random, and

we consider the typical performance over an instantiation of the problem rather than the worst possible case, no learning algorithms are known to achieve the goal in a time which goes as a polynomial of the number of synapses, including models which make use of meta-plasticity.

Recently, 'message passing' algorithms have been devised that solve efficiently non-trivial random instances of NP-complete optimization problems, like e.g. K-satisfiability or graph coloring [**19, 20, 1, 7**]. One such algorithm, Belief Propagation (BP), has been applied to the binary perceptron problem and has been shown to be able to find efficiently synaptic weight vectors that solve the classification problem for a number of patterns close to the maximal capacity (above 0.7 bits per synapse)[**8**]. However, this algorithm has a number of biologically unrealistic features (e.g. memory stored in several analog variables).

Here, we present a novel algorithm, called SBPI, that is inspired from the BP algorithm but is modified in order to make it simpler and biologically realistic, while keeping very good performances, both in terms of learning time and of information storage capacity. This algorithm requires meta-plasiticity, and introduces also a new learning rule, directly inherited from the BP algorithm, which also proves to be able to boost the performance of other algorithms. We provide evidence about the qualitatively superior performance of SBPI with respect to other well known algorithms both using extensive computer simulations, in the case of learning a set of pattern classifications, and by an analytical mean field study, in the case of learning a rule from a teacher device.

UNIVERSITÀ DEGLI STUDI DI TORINO

Dipartimento di Chimica Generale ed Organica Applicata e
Dipartimento di Scienze Oncologiche

Dottorato di Ricerca in
SISTEMI COMPLESSI APPLICATI ALLA BIOLOGIA POST-GENOMICA

XX CICLO

TITOLO DELLA TESI:

Protein-Ligand and Protein-Protein Interaction Studies: Development of a Spectroscopy-based Approach

TESI PRESENTATA DA:
Nadia Barbero

TUTOR:
Prof. Federico Bussolino

COORDINATORE DEL DOTTORATO:
Prof. Federico Bussolino

Anni Accademici: **2004/2007**

Settore Scientifico-Disciplinare: **CHIM/06**

Abstract

A complete signaling cascade was reconstructed by mathematical description of the INTEGRIN/VEGFR-induced pathways and the attention was focused on a useful group of proteins to work on. Some proteins, which do not belong to the pathway, were chosen to be used as a test for the development of a method for interaction studies. Four different interactions were used in order to explore four different kinds of interactions: protein-ligand, protein-antibody, protein-peptide and protein-protein interactions.

The binding of fluorescein sodium salt with different bovine serum albumins (BSA) was investigated by steady-state and stopped-flow fluorescence. This interaction was chosen for a preliminary study for protein-ligand interactions because of BSA low cost and availability. The dissociation and association rate constants (k_{on} and k_{off}) were determined from the kinetic studies while the dissociation and association binding constants (K_d and K_a) were determined both by the quenching of the fluorescence of BSA in the presence of fluorescein and from stopped-flow measurements from the k_{on}/k_{off} ratio. This work also reports the distance between tryptophan and bound fluorescein based on Förster's energy transfer theory and a thermodynamic study of the mode of interaction which is important for confirming binding modes.

The interaction between GST with his antibody α-GST (B14) was studied by fluorescence anisotropy, after GST bioconjugation with fluorescein-5-maleimide, leading to the K_d and K_a determination which have resulted in agreement with the data found in literature referring to protein-antibody interaction.

GST-Tat86 interaction with two small peptides (CT319 and V2) having similar aa sequence and the same biological activity was studied by steady-state fluorescence exploiting the intrinsic Trp residues fluorescence. The results show that the obtained binding constants differ in the same set of interactions. One set of results is in agreement with the K_d obtained from BIAcore studies and reported by Marchiò *et al.* (*Blood* **2005,** *105,* 2802-2811) for a very similar interaction. The variation of the obtained binding constants can be due to the presence of many tryptophan residues each with a different environment. The reaction was then followed by steady-state and stopped-flow fluorescence after the peptide bioconjugation.

Finally, the interaction between two proteins belonging to the above-mentioned pathway was studied. The interaction between MEK-ERK was followed by stopped-flow fluorescence after ERK bioconjugation with fluorescein-5-maleimide. The second-order rate constant k_{on} is $2.84 \cdot 10^8$ M^{-1} sec^{-1} while the k_{off} has a small negative value. This shows that the interaction is nearly completely shifted toward the complex formation. The binding constant

obtained is in agreement with the few data available in literature referring to this interaction obtained either from simulation/prediction analysis or different techniques.

Therefore, this work has provided evidences of the possibility of studying protein-ligand interactions by spectroscopic methods but has also outlined all the difficulties and the limits of the reported techniques. In particular it should be stressed the need of high quantity of high purity proteins or, more generally, biomolecules.

UNIVERSITY OF TURIN

DOTTORATO DI RICERCA IN SISTEMI COMPLESSI APPLICATI ALLA BIOLOGIA POST-GENOMICA

COMPLEX SYSTEMS IN POST-GENOMIC BIOLOGY

CYCLE XX

TITLE:

GENOME WIDE ANALYSIS OF ESTROGEN RESPONSE ELEMENT (ERE) DISTRIBUTION: FUNCTIONAL ANALYSIS OF ESTROGEN-INDUCED GENE REPRESSION

CANDIDATE:
Maria Cardamone

TUTOR:
Prof. Michele De Bortoli

PhD COORDINATOR:
Prof. Federico Bussolino

ACCADEMIC YEARS: 2004/2007 **SCIENTIFIC FIELD: BIO/11**

ABSTRACT

Estrogens play an essential role in both physiological development and breast cancer progression, but the gene networks and pathways by which estrogenic hormones regulate these events are only partially understood.

In the last few years several approaches to computational prediction of functional binding sites have been developed. They are all based on one pattern matching that usually is the representation as a matrix of acceptable nucleotides at each position of the known binding sites for a given protein. Following this criteria and using the data generated from chromatin immunoprecipitation (ChIP) on chip experiment (Kwon et al., 2007) (Carroll et al., 2006) we built a new ERE weighted alignment matrix (ERE-m). We used this matrix in a pattern discovery algorithm to perform a genome-wide scanning for putative estrogen responsive genes. To eliminate false positive motif we use the phylogenetic sequence conservation. We focused our attention on down-regulated ERE-containing genes and we validated the *in silico* analysis by ChIP and expression studies.

An interesting gene in this group is CDH-1, because it encodes for E-Cadherin, a trans-membrane protein important for cell-cell adhesion and involved in Epithelial-Mesenchimal Transition (EMT), a natural event during development that plays a key role in tumor progression. We demonstrated that ERα is recruited at the E-cadherin promoter even in the absence of estrogen stimulation, in breast cancer cells. Moreover we demonstrated that in absence of estrogen stimulation ERα is required to maintain the basal level of CDH-1 expression, while in presence of the ligand it becomes a repressor. Our data suggest a possible new role for ERα as ligand-independent activator that can be essential for the determination of epithelia morphology.

Our results show that the same factor (ER) bound to the same sequence (ERE) can evoke either activation or repression at different gene contexts. This may be explained by the hypothesis that transcriptional complexes with distinct composition exist in the nuclei, taking care of the transcription of distinct subsets of genes, in response to the same stimulus. For this reason, I joined the laboratory of M.G. Rosenfeld, who was examining the possibility that genes with a common mode of regulation in response to stimuli can share the same transcriptional machinery. Results of this study demonstrated that ligand induces rapid interchromosomal interactions among subsets of

estrogen receptor α-bound transcription units, with a dramatic reorganization of nuclear territories requiring nuclear actin/myosin-1 transport machinery, dynein light chain 1, and a specific subset of transcriptional coactivators and chromatin remodeling complexes. We establish a molecular mechanism by which the hormone-induced interchromosomal interactions serving to achieve enhanced, coordinated transcription and RNA splicing for nuclear receptor target genes.

UNIVERSITA' DEGLI STUDI DI TORINO

DIPARTIMENTO DI SCIENZE ONCOLOGICHE

Dottorato di ricerca in

SISTEMI COMPLESSI APPLICATI ALLA BIOLOGIA POST-GENOMICA

Ciclo XX

Titolo:

A novel Rab5-based signalling pathway participates in centrosome cohesion.

Tesi presentata da:
Valentina Margaria

Tutor:
Prof. Federico Bussolino

Coordinatore del Dottorato:
Prof. Federico Bussolino

Anni accademici: 2004/2005 - 2005/2006 - 2006/2007

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA:
BIO/10

Abstract

The centrosome is a non-membranous organelle that acts primarily as a microtubule organising centre. During interphase, centrosomes organise the microtubule network responsible for vesicular transport, for cell shape and polarity. At mitosis, centrosomes direct the formation of the mitotic spindle and ensure proper separation of replicated chromosomes.

The centrosome is composed of two paired orthogonal centrioles that replicate at the G1/S transition giving rise to two centrosomes that are held together until G2 when they separate migrating to the opposite poles of the cell.

The control of centrosome cohesion is an important aspect of cell division, since its deregulation can affect spindle assembly. Not surprisingly, therefore, deregulation of centrosome division and dynamics is thought to play a major role in genomic instability associated with tumorigenesis.

Here we report the identification of a novel Rab5-dependent pathway participating in centrosomes separation.

Rab5 is a small GTPase involved in the control of intracellular trafficking, homeostasis of the endosomal compartment and actin cytoskeleton remodelling. The activity of Rab5 is tightly regulated by GDP/GTP exchange factors (GEFs), like Rabex-5, and GTPase activating protein (GAPs), such as RN-tre.

We found that Rab5, RN-tre and Rabex-5 localise at the centrosome in human cells. Moreover increased Rab5 activity caused loss of centrosome cohesion suggesting a role for this GTPase in centrosome function. Indeed, reduction of the Rab5 activity inhibited centrosome separation during G2 and decreased the distance between the spindle poles at mitosis.

The molecular mechanism appears to involve the kinesin motor protein KIF3A. KIF3A binds to RN-tre and its depletion prevented the loss of centrosome cohesion caused by excess of active-Rab5.

More importantly, KIF3A silencing phenocopies the effects of Rab5 depletion, inhibiting centrosome separation and causing defective spindles.

Thus KIF3A has the characteristics of a Rab5-downstream effector and participates in the separation of duplicated centrosomes.

UNIVERSITA' DEGLI STUDI DI TORINO

DIPARTIMENTO DI SCIENZE BIOMEDICHE

DOTTORATO DI RICERCA IN: "Complex systems applied to post-genomic biology"

CICLO: XX

Functional screening for genes conferring anchorage-independence to human immortalized breast cells

TESI PRESENTATA DA:
Mira Alessia

TUTOR:
Prof. Medico Enzo

COORDINATORE DEL CICLO
Prof. Bussolino Federico

ANNI ACCADEMICI: 2004-2007

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA*: BIO/17

Abstract

Cells sense their location through specific interactions with the extracellular matrix (ECM) and neighbouring cells. As well as providing positional and mechanical cues for the organization of cells in tissues, the ECM is able to control fundamental cellular fates including growth, differentiation, survival and movement. Integrin-mediated regulation of cell survival is thought to contribute to maintenance of tissue homeostasis by ensuring that cells remain in their proper tissue environment. Disruption of this connection leads to a specific type of apoptosis known as anoikis in most non transformed cell types. Conversely, tumor cells display significant resistance to anoikis, as most tumor-derived epithelial cells survive in the absence of adhesion. Anchorage independent growth is a critical step in the metastatic transformation of a tumor since malignant cells, once they begin to metastasize, acquire properties rendering them able to detach from the primary tumor and to escape apoptotic mechanisms. Many oncogenes mimic integrin-dependent signals, thus allowing cancer cells to survive under conditions in which normal cells undergo apoptosis. The identification and characterization of novel genes or proteins promoting anchorage-independent growth, and thereby potentially driving a malignant phenotype, will likely provide candidate targets for innovative anticancer therapies.

Functional genomics aims to assign functional information to each of the genes in an organism in a high-throughput, systematic manner. Several technologies have recently become available to identify gene function in mammalian cells through “gain” or “loss” of function screening; however actual genetic screening are hampered by difficulties in the cloning and identification of integrated ORFs giving a particular phenotype, rendering them still high-demanding and labour-intensive. Here we conceived a high-throughput gene functional screening named “Xenoarray analysis” in which standard gene expression arrays are used for tracing the abundance of exogenous cDNAs derived from the library, before and after a selection. The screening is carried out by applying a selective pressure on the transduced cell population, so that cells transduced with genes conferring a selective advantage are enriched. Comparing microarray signal intensities for the exogenous genes before and after selection allows simultaneous detection of all advantageous ones. This

technology has the advantage to overcome previous problems related to genetic screens by detecting selected enriched ORFs in one single experiment. In this study, we transduced MCF10A cells (spontaneously immortalized human breast epithelial cells), with a retroviral mouse testis library and applied a selection by growing cells in the absence of anchorage. By DNA microarray analysis, we were able to trace the abundance of the enriched ORFs conferring a growth advantage in the absence of anchorage. In two independent infections and selections, the most reproducible enriched gene resulted to be GAB2. Gab2 cDNA transduction in MCF10A cells enabled the validation of its role in anchorage-independent growth.

Università Degli Studi Di Torino

Facoltà di Scienze Matematiche Fisiche e Naturali

Dipartimento di Fisica Teorica

Dottorato di Ricerca in Sistemi Complessi Applicati alla Biologia Post-Genomica

CICLO XX

PARALOGOUS ALIGNMENTS AS A TOOL TO INVESTIGATE GENOMIC INFORMATION

TESI PRESENTATA DA:
Dott. Ivan Molineris

TUTOR:
Prof. M. Caselle

COORDINATORE DEL CICLO:
Prof. F. Bussolino

RELATORI ESTERNI:
Prof. G. Valle
Prof. C. Herrmann

Anni Accademici: 2005 - 2007
SETTORE SCIENTIFICO-DISCIPLINARE DI APPARTENENZA: FIS-02

Abstract

In the last few years the amount of information about genomes, especially accurate complete sequences, has been exponentially increasing. Despite this abundance of information the interpretation in biological significant terms of the entire genomic sequence of an organism remains a challenge of the post-genomic scientific era.

In this study we propose paralogous alignments (i.e. the alignments of a genome with itself) as a tool to extract meaningful information from raw genomic sequences. We computed a complete database of such alignments for a few organisms and we developed a set of software tools to mine, collect, visualize and integrate these data with the present knowledge about genomic sequences.

As a first result we were able to identify previously unknown genes (chapter 2).

As a second step we adopted a more abstract perspective which was inspired by two considerations: on one hand many known languages are structured so that the used words are only few of the possible combinations of characters; on the other hand, a word is often present many times in the same text. Therefore we searched for sequences of nucleotides occurring many times in the genomes using paralogous alignments and managing them with graph theory concepts. Within the "genomic words" that we found there are many well known sequences, such as protein domains, but also previously uncharacterized sequences (chapter 3).

Alongside this principal project, which is centred on a genomic level, we investigated other ways to extract meaningful information from large sets of publicly available biological data. We developed a strategy able to handle gene expression profiles in order to infer interactions among genes at proteomic level starting from data at the transcriptomic level (appendix A).

Universitá di Roma Tor Vergata

Facoltá di Scienze Matematiche Fisiche e Naturali

Dipartimento di Biologia

Dottorato di Ricerca in Sistemi Complessi Applicati alla Biologia Post-Genomica

CICLO XX

COMPUTATIONAL FRAMEWORKS FOR WIRING HUMAN AND YEAST PROTEOMES

TESI PRESENTATA DA:
Dott.ssa Maria Persico

TUTOR:
Prof. R. Calogero

COORDINATORE DEL CICLO: Prof. F. Bussolino
RELATORE ESTERNO: Prof. G. Cesareni

Anni Accademici: 2004 - 2007
SETTORE SCIENTIFICO-DISCIPLINARE DI APPARTENENZA: BIOL-

Abstract

One of the major challenges of modern system biology is to decipher how the information for the life processes is encoded in the protein networks of a complex organism. Interactions among proteins serve as an important basis for the biological complexity of higher organisms. In recent years, there have been several large scale efforts to map protein interactions on model organisms.

In this thesis, I address the problem of how to wire the proteomes, an important question, central in systems biology; in human (*H. sapiens*), the difficulties in generating protein interaction data have stimulated the development of sequence based prediction frameworks, i.e. co-evolutionary information of the interacting partners and interologs; following this trend, some pipelines were developed aimed to transfer interaction data from model organism to human (the HomoMINT interactome) and to looking for correlations in the distance matrices representing the trees of the ortholog groups to which the human reference proteins under analysis belong. A third method based on the identification of co-evolving residues displaying statistically significant patterns of co-evolution, as measured by mutual information metric was tested; in yeast (*S. cerevisiae*) , the questions related to the wiring problem remain still unanswred in spite of the abundance of protein interaction data from high-throughput experiments. Unfortunately, these large-scale studies show embarrassing discrepancies in their results and coverage. The recent completion of a comprehensive literature curation effort, have made available an interesting new reference set and stimulated building of a simple logistic regression model on wiring of the yeast proteome, based on the definition of some predictors of functional relationships for the pairs of interacting proteins in the reference set: the probability of sharing a path on the Gene Ontology trees, the degree of correlated evolution, the degree of co-expression and co-abundance. Moreover, the value distributions for the analysed genomic features differ respect the distributions of the same predictors computed in previously defined null models (i.e. artificial protein networks).

The model was evaluated by standard criteria and ROC curve analysis. The complete frameworks were implemented in a suite of R - PERL programs.

“Detection of Pharmacophores associated to drugs toxicity”

By

SAID M. ZAMIT

Under supervision of:

Prof. Raffaele Calogero

Abstract

The ability to reliably predict *in vivo* toxicity through *in vitro* models is increasing. The use of human cultured cell lines seems to be especially promising both for acute and chronic toxicity evaluation. However the techniques currently used, some of which based on the measurement of protein and ATP content and cell morphology, suffer of the restriction of this simplified end-point data evaluation which proves to be inadequate for prediction of organ-specific toxicity and toxicity of substances that do not induce cell death.

The goal of computational toxicity prediction is to describe possible relationships between chemical properties of the drug as well as biological and toxicological process or mechanism. In many cases the important points of interaction between a drug and its target can be represented by a 3D arrangement of a small number of atoms. Such a group of atoms is called pharmacophore. A pharmacophore can be used to search 3D databases of drugs and compounds sharing the pharmacophore can belong to different chemical classes.

In this thesis I'm searching for correlation between drug toxicity and pharmacophores using a 3D library of compounds, and their toxicity index on different cell lines. Here, with pharmacophore (toxiphore) searching I'm interested to detect local similarity, i.e. based on a limited number of atoms (e.g. 3,4 atoms) within high toxic compounds. My hypothesis is that such similarities could be dealt with their high toxicity. The final aim of this study is the definition of a Drug Toxicological Index (DTI). This index should be able to predict the

toxicity strength of new compounds before they are going into practical experimentation. DTI will be defined upon identification of pharmacophores (toxicophores) associated to toxicity, and the most important part of the study is finding the toxicophores related with toxicity .

This work is based on meta-analysis of public available data, The used databases are NCI DIS 3D database (http://129.43.27.140/), and Corina dataset (http://129.43.27.140/ncidb2/download) which are a collection of 3D structures for over 500,000 drugs, each which was built and is maintained by the Developmental Therapeutics Program "DTP", Division of Cancer Treatment, National Cancer Institute, Rockville ,MD. At NCI 3,000 compounds per year are screened for their potential anticancer activity. The DTP Human Tumor Cell Line Screen has checked tens of thousands of screened compounds for evidence of the ability to inhibit the growth of human tumor cell lines. This screen utilizes 60 different human tumor cell lines, representing leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney.

Screened drugs are saved in MOL format, and I have converted them into a tabular form and loaded into MYSQL relational database. I stored the structure information together with toxicity index and I used this data to search for drugs that share three atoms pharmacophore.

To detect "high toxic" pharmacophores, I collected the compounds that shows high toxicity index over all cell lines, then I extracted all possible toxicophores. Those toxicophores were then scanned across all very low toxic compounds and I found that these suspected toxicophores were under represented.

Out of a total of twenty six toxicophores found, six of them are found in compounds with toxicity index greater than 6, the other in compounds with toxicity index between 5 And 6.