UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Lilian Hernández Alvarez

Identification and Characterization of Cruzain Allosteric Inhibitors: A Computer-Aided Approach

São José do Rio Preto
2017

Lilian Hernández Alvarez

Identification and Characterization of Cruzain Allosteric Inhibitors: A Computer-Aided Approach

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Orientador: Prof. Dr. Pedro Geraldo Pascutti

São José do Rio Preto

2017

Hernández Alvarez, Lilian.

Identification and characterization of cruzain allosteric inhibitors : a computer-aided approach / Lilian Hernández Alvarez. -- São José do Rio Preto, 2017

102 f. : il., tabs.

Orientador: Pedro Geraldo Pascutti

Dissertação (mestrado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Biologia molecular. 2. Biofísica. 3. Bioinformática. 4. Trypanosoma cruzi. 5. Fármacos. 6. Cisteína proteases. 7. Inibidores enzimáticos. I. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. II. Título.

CDU – 577.3

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Lilian Hernández Alvarez

Identification and Characterization of Cruzain Allosteric Inhibitors: A Computer-Aided Approach

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Biofísica Molecular, junto ao Programa de Pós-Graduação em Biofísica Molecular, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Comissão Examinadora

Prof. Dr. Pedro Geraldo Pascutti
UFRJ – Rio de Janeiro
Orientador

Dr. Ícaro Putinhon Caruso
UNESP – São José do Rio Preto

Prof. Dr. Fernando Luís Barroso da Silva
USP – Ribeirão Preto

São José do Rio Preto
29 de setembro de 2017

Inscription

A todos aquellos que me inspiran día a día desde
la distancia…

Acknowledgments

This project would not have been possible without the support of many people. Many thanks to my adviser, Pedro Geraldo Pascutti, who brought me here and put all his confidence in me. Without him nothing would has been possible. Also thanks to my coworker Diego H. Gomes, who offered guidance and support. Thanks to the CAPES for awarding me this fellowship, providing me with the financial means to complete this project. And finally, thanks to my husband, daughter, and numerous friends who accompanied me in this long process, always offering support and love.

“My mission in life is not merely to survive, but to thrive; and to do so with some passion, some compassion, some humor, and some style."

— Maya Angelou

RESUMO

Trypanosoma cruzi é o agente causal da doença de Chagas, uma infecção negligenciada que afeta milhões de pessoas nas regiões tropicais. A maioria dos fármacos empregados no tratamento desta doença são altamente tóxicos e geram resistência. Na atualidade, o descobrimento de inibidores alostéricos é um tópico emergente dentro da área de desenho computacional de fármacos, pois promove a acessibilidade a medicamentos mais seletivos e menos tóxicos. Neste trabalho foi desenvolvida uma estratégia para a descoberta computacional de inibidores alostéricos a qual foi aplicada à cruzaína, a principal cisteíno protease do *T. cruzi*. A caracterização molecular da forma livre e ligada da cruzaína foi investigada através do ancoramento molecular, simulações de dinâmica molecular, cálculos de energia livre de ligação e construção de redes de interações entre resíduos. A partir da análise baseada na geometria das estruturas geradas na dinâmica molecular, foram detectados dois potenciais sítios alostéricos na cruzaína. Os resultados sugerem a existência de diferentes mecanismos de regulação exercidos pela ligação de inibidores diferentes no mesmo sítio alostérico. Além disso, foram identificados os resíduos que estabelecem os caminhos de transmissão de informação entre um dos sítios alostéricos identificado e o sítio ativo da enzima. O presente estudo é a primeira aproximação de desenho de inibidores alostéricos da cruzaína e serve para futuras intervenções farmacológicas. Esses resultados constituem uma base para o desenho de inibidores específicos de cisteíno proteases homólogas da papaína.

Palavras-chave: cruzaína, alosteria, triagem virtual, energia livre de ligação, correlação generalizada, redes de proteínas, comunidades, caminho subótimo.

ABSTRACT

Trypanosoma cruzi is the causative agent of Chagas disease, a neglected infection affecting millions of people in tropical regions. There are several chemotherapeutic agents for the treatment of this disease, but most of them are highly toxic and generate resistance. Currently, the development of allosteric inhibitors constitutes a promising research field, since it may improve the accessibility to more selective and less toxic medicines. To date, the allosteric drugs prediction is a state-of-the-art topic in rational structure-based computational design. In this work a simulation strategy was developed for computational discovery of allosteric inhibitors, and it was applied for or cruzain, a promising target and the major cysteine protease of T. cruzi. Molecular dynamics simulations, binding free energy calculations and network-based modelling of residue interactions were combined to characterize and compare molecular distinctive features of the apo form and the cruzain-allosteric inhibitor complexes. By using geometry-based detection on trajectory snapshots we determined the existence of two main allosteric sites suitable for drug targeting. The results suggest dissimilar mechanism exerted by the same allosteric site when binding different potential allosteric inhibitors. Finally, we identified the residues involved in suboptimal paths linking the identified site and the orthosteric site. The present study constitutes the first approximation for designing cruzain allosteric inhibitors and may serve for future pharmacological intervention. These findings are particularly relevant for the design of allosteric modulators of papain-like cysteine proteases.

Keywords: *cruzain, allostery, virtual screening, binding free energy, generalize correlation, protein network, communities, suboptimal pathway*.

Figure list

Table List

Nomenclature and abbreviations

APBS: Adaptive Poisson-Boltzmann Solver

CC: Cross-correlation

CDC: Centers for Diseases Control

CP: Coordination Propensity

DNA: Deoxyribonucleic Acid

ED: Essential Dynamics

EM: Energy Minimization

FDA: Agency of Food and Drug Administration

FEP: Free Energy Perturbation

GAFF: General AMBER Force Field

GAGs: Glycosaminoglycans

GC: Generalized Correlation

GPCR: G Protein-Coupled Receptors

HAT: Human African Trypanosomiasis

HCatB: Human Cathepsin B

HCatK: Human Cathepsin K

HCatS: Human Cathepsin S

IC50: The half-maximal inhibitory concentration

LIE: Linear Interaction Energy

Lmcps: *Leishmania* cathepsins

MD: Molecular Dynamics

MM-GBSA: Molecular Mechanics/Generalized Born Surface Area

MM-PBSA: Molecular Mechanics/Poisson Boltzmann Surface Area

NF-kB: Nuclear Factor *kappa B*

NMR: Nuclear Magnetic Resonance Spectroscopy

NPT: Constant-pressure ensemble

NTD: Neglected Tropical Diseases

NVT: Constant-volume ensemble

Orthosteric site: Describes the primary site of a receptor (protein)

Orthosteric drug: Drug binding to the orthosteric site

PCA: Principal Component Analysis

PDB: Protein Data Bank

PME: Particle Mesh Ewald method

POVME: POcket Volume MEasurer

RESP: Restricted Electrostatic Potential

RMSD: Root-Mean-Square Deviation

RMSF: Root-Mean-Square Fluctuation

SCA: Statistical Coupling Analysis

TGF-β: Transforming grow factor *beta*

TI: Thermodynamic Integration

VS: Virtual Screening

WHO: World Health Organization

WISP: Weighted Implementation of Suboptimal Paths

ΔG_{bind} :Binding Free energy

ΔG_{eff} : Effective Free Energy

ΔG_{res} :Per-residue Free Energy Contribution

Table of contents

1. Introduction

Cruzain is the major papain-like cysteine protease of *Trypanosoma cruzi*, the protozoan responsible for Chagas disease. This enzyme is indispensable for the survival and propagation of this parasite and, therefore, is considered as a potential drug target for this disease control (McGrath,Eakin *et al.* 1995;Mark W. Robinson PhD 2011;Sajid,Robertson *et al.* 2011). Toxicity and resistance of the available chemotherapy (Ribeiro,Sevcsik *et al.* 2009;Wilkinson and Kelly 2009;Pena,Pilar Manzano *et al.* 2015), fueled the pursuit of alternative chemotherapeutic agents, leading to the discovery of several cruzain modulators. A tangible evidence of this, is the existence of twenty-five crystal structures of cruzain in complex with several competitive inhibitors on the Protein Data Bank (PDB) (McGrath,Eakin *et al.* 1995;Mott,Ferreira *et al.* 2010;Rogers,Keranen *et al.* 2012;Martinez-Mayorga,Byler *et al.* 2015). On the other hand, several experimental studies and computational predictions have been performed to characterize the cruzain binding site and specificity, facilitating the design of active-site directed drugs (Trossini,Guido *et al.* 2009;Durrant,Keranen *et al.* 2010;Ferreira,Simeonov *et al.* 2010;Rogers,Keranen *et al.* 2012;Martinez-Mayorga,Byler *et al.* 2015). However, the competitive inhibitors of cruzain have demonstrated to be toxic in several clinical trials because the high identity of this enzyme with its host homologues. This fact has encouraged the search for new scaffolds of cruzain competitive inhibitor as well as different strategies for the enzyme inhibition and ultimately novel therapeutic targets.

Allostery phenomenon is an inherent characteristic of most macromolecules (Dokholyan 2016;Nussinov 2016). In recent decades, the design of allosteric drugs has emerged as a promising research field in disease control (Nussinov and Tsai 2013;Lu,Li *et al.* 2014;Liu and Nussinov 2016). Allosteric modulators offer many advantages that make them appropriate as drug candidates. A major benefit of these strategies, compared to those that perturb the active site directly, is that they offer a noninvasive or more specific protein control. Allosteric regulators do not interfere with modulators of active site and do not obstruct its vicinity (Tsai and Nussinov 2014;De Vivo,Masetti *et al.* 2016;Dokholyan 2016).

Several findings pointing to the existence of allosteric regulation in the superfamily of papain-like cysteine proteases (Costa,dos Reis *et al.* 2012;Judice,Manfredi *et al.* 2013;Jilkova,Horn *et al.* 2014;Novinec,Lenarcic *et al.* 2014). One example is the work of Novinec *et al.* where several allosteric sites of human cathepsin K (HCatK) and the sectors which mediate allosteric communication in this protein family were identified employing the statistical coupling analysis (SCA) method (Novinec,Korenc *et al.* 2014). Moreover, in that work and in a more recent study, the crystallization of HCatK with allosteric inhibitors, positioned in one of a previously-predicted site, was reported (PDBID: 5J94 and 5JA7) (Novinec,Korenc *et al.* 2014;Novinec,Lenarcic *et al.* 2014;Novinec,Rebernik *et al.* 2016). On the other hand, there are other cases of allosteric modulation in several parasitic cysteine proteases such as Falcipain 2 from *Plasmodium falciparum*, which are allosterically inhibited by heme and suramin compounds (Marques,Esser *et al.* 2013;Marques,Gomes *et al.* 2015).

Interestingly, the role of glycosaminoglycans (GAGs) in allosteric modulation of papain-like cysteine proteases has been widely studied in the past decades. Proteases such as human cathepsins S (HCatS) and B (HCatB), brucipain, cathepsin L of *Leishmania mexicana* and even cruzain present an allosteric modulation by GAGs (Almeida,Nantes *et al.* 1999;Almeida,Nantes *et al.* 2001;Lima,Almeida *et al.* 2002;Li,Yasuda *et al.* 2004;Costa,Batista *et al.* 2010;Costa,dos Reis *et al.* 2012;Judice,Manfredi *et al.* 2013). However, no other biomolecules or chemical compounds have been reported as allosteric modulators of cruzain. Indeed, only one *in silico* characterization of an adjacent subsite to the cruzain orthosteric site was made by Durrant *et al.* (Durrant,Keranen *et al.* 2010).

In recent years, the computer-aided design of allosteric inhibitors has been widely used in the field of rational drug discovery (Dalafave and Dalafave ;Verkhivker,Dixit *et al.* 2009;Novinec,Lenarcic *et al.* 2014;Rastelli,Anighoro *et al.* 2014). The application of techniques such as SCA, molecular docking, molecular dynamics (MD) simulations and allosteric networks has emerged as a valuable complement to experimental methods for the study of allostery (Hertig,Latorraca *et al.* 2016;Nussinov 2016;Papaleo,Saladino *et al.* 2016;Ribeiro and Ortiz 2016;Wagner,Lee *et al.* 2016). These approaches have also allowed the prediction of allosteric sites and the quantification of protein and/or ligand motions in full atomic detail, describing the molecules behavior at high resolution and with the induction of controlled perturbations. (Stanley and De Fabritiis 2015;Zhao and Caflisch 2015;De Vivo,Masetti *et al.* 2016;Hernandez-Rodriguez,Rosales-Hernandez *et al.* 2016).

In this sense, the **main objective** of this work is the following:

To identify allosteric sites and selective allosteric inhibitors of cruzain through *in silico* approaches

In order to accomplish this general objective we will follow the **specific objectives**:

- To select the five crystal structures of cruzain from PDB database.
- To identify putative allosteric sites of this enzyme through structural alignment with HCatK.
- To search for three representative structures of each enzyme through a combination of MD simulations and clustering analysis, in order to perform ensemble virtual screening (VS) experiments with compounds from ZINC database against the putative allosteric sites.
- To re-rank the compounds selected from VS and to refine their binding modes through MD simulations of the enzyme-ligand complexes, and to perform free energy calculations with Molecular Mechanics/Generalized Born Surface Area (MM-GBSA) method based on NPT ensembles extracted from the MD simulations.
- To map the interactions of the selected ligands with the enzyme through per-residue free energy decomposition employing the MM-GBSA method.
- To detect conformational changes in trajectories of apo and holo forms of cruzain evaluating some parameters such as changes in interatomic distances on orthosteric site, coordination propensity and principal component analysis.
- To predict the molecular basis of the cruzain allosteric mechanism by performing cross-correlation analysis of residue-residue distances of apo and holo MD simulations and prediction of the communication pathways between the allosteric and the active sites.

2. Literature review

2.1. Chagas disease as a Neglected Tropical Disease

2.1.1. General aspects of Neglected Tropical Diseases

The concept of neglected tropical diseases (NTDs) emerged more than a decade ago and has been recognized as a valid way to categorize diseases that affect the poorest countries. The majority of individuals and communities in affected regions have far less access to the resources necessary to address the social determinants of NTDs. They also may live in poor sanitary conditions, have inadequate nutrition, and lack access to necessary public health and health care systems for treatment, despite many of these diseases being preventable and/or treatable through specific low-cost interventions (Mackey,Liang *et al.* 2014;Hotez and Bundy 2017).

World Health Organization (WHO) has specifically identified 17 core NTDs: dengue, rabies, trachoma, buruli ulcer, endemic treponematoses, leprosy, Chagas disease, Human African Trypanosomiasis (HAT), leishmaniasis, taeniasis/cysticercosis, dracunculiasis, echinococcosis, food-borne trematodiases, lymphatic filariasis, onchocerciasis, schistosomiasis, and soil-transmitted helminthiases. NTDs are comprised primarily of viral, protozoan, helminthial, and bacterial infections (Mackey,Liang *et al.* 2014).

Reducing the burden of NTDs is one of the key strategic targets advanced by the sustainable development goals. Despite the unprecedented effort deployed for NTD elimination in the past decade, their control, mainly through drug administration, remains particularly challenging: persistent poverty and repeated exposure to pathogens embedded in the environment limit the efficacy of strategies focused exclusively on human treatment or medical care. Efforts to protect the health of these populations have been insufficient, with the global focus to identify and prioritize NTDs by the international community only “reemerging” in the last decade following efforts by leading NTD researchers and advocates (Hotez and Bundy 2017).

2.1.2. Trypanosomiasis

Kinetoplastids are a group of protozoans conformed by both: (*i*) obligated parasites and (*ii*) free-living organism. The principal kinetoplastids with health and economic importance are trypanosomatids, which belong to obligated parasites subgroup.

Currently, *Trypanosoma* and *Leishmania* species affect millions of people and animals throughout the world. In sub-Saharan Africa *Trypanosoma brucei gambiense* and *Trypanosoma brucei rhodesiense* are the causative agents of human African trypanosomiasis (HAT or sleeping sickness), while *T. brucei*, *T. congolense* and *T. vivax* are responsible of 'Nagana' disease, which affects cattle. In addition, South and Central America are affected by *T. cruzi*, the causative agent of Chagas disease in humans, which has become in the principal heart illness of this region (Mark W. Robinson PhD 2011).

The American trypanosomiasis was discovered in 1909 by scientist Carlos Chagas, which constituted one of the most successfully discoveries in the history of tropical medicine. However, the natural history of trypanosomiasis began approximately 7-10 million years ago when the predecessor of *T. cruzi* was probably introduced in South American via bats and was preserved as an enzootic disease between wild animals. Currently, this kind of propagation persists in regions such as Amazonia. Over the last 200-300 years, the endemic Chagas disease was established as a zoonosis through deforestation for agriculture and cattle rearing. Moreover, triatomines adaptation to houses and to humans and domestic animals as food sources was critical for disease dissemination (Coura and Borges-Pereira 2010;Steverding 2014).

To date, Chagas disease is responsible of 50 000 deaths per year. In addition, the WHO estimates that 7-10 million individuals are infected in worldwide and approximately 90-100 million are exposed to the infection (Ferreira and Andricopulo 2017). This disease has an initial or acute phase, with notable parasitemia detected in direct blood tests. In most cases, it is asymptomatic, but symptomatic cases can show entry point signs (inoculation chagoma or Romaña's sign), fever, generalized adenopathy, edema, hepatosplenomegaly, myocarditis and meningoencephalitis in severe cases. The acute phase is followed by a chronic phase, which usually is presented as an indeterminate form (normal electrocardiogram and normal radiographs on the heart, esophagus and colon). The chronic cardiac form is the most significant clinical manifestation of Chagas disease because of its frequency and severity. The chronic cardiomyopathy generally appears between the second and fourth decades of life, generally 5-15 years after initial infection. The signs and symptoms of cardiomyopathy caused by Chagas disease are the arrhythmia, heart failure, atrioventricular and branch blocks and thromboembolism. All these symptoms are mainly originated by the action of immune system, which produces inflammatory lesions, cell destruction and fibrosis. There is great regional variation in the morbidity caused by Chagas disease, where severe cardiac or digestive forms may occur in 10-50%, and indeterminate forms in the remaining asymptomatic cases (Coura and Borges-Pereira 2010;Coura,Vinas *et al.* 2014).

2.1.3. Life cycle of *T. cruzi*

The biological cycle of *T. cruzi* occurs through a continuous transformation between the states that colonize mammals (including humans) and those that colonize insects. Transmission of Chagas disease is mediated by hematophagous of *Triatoma* (reduviid or vinchuca) and *Rhodnius* genera, also known as kissing/killer bug (Figure 1) (Noireau,Diosque *et al.* 2009;Coura and Borges-Pereira 2010).

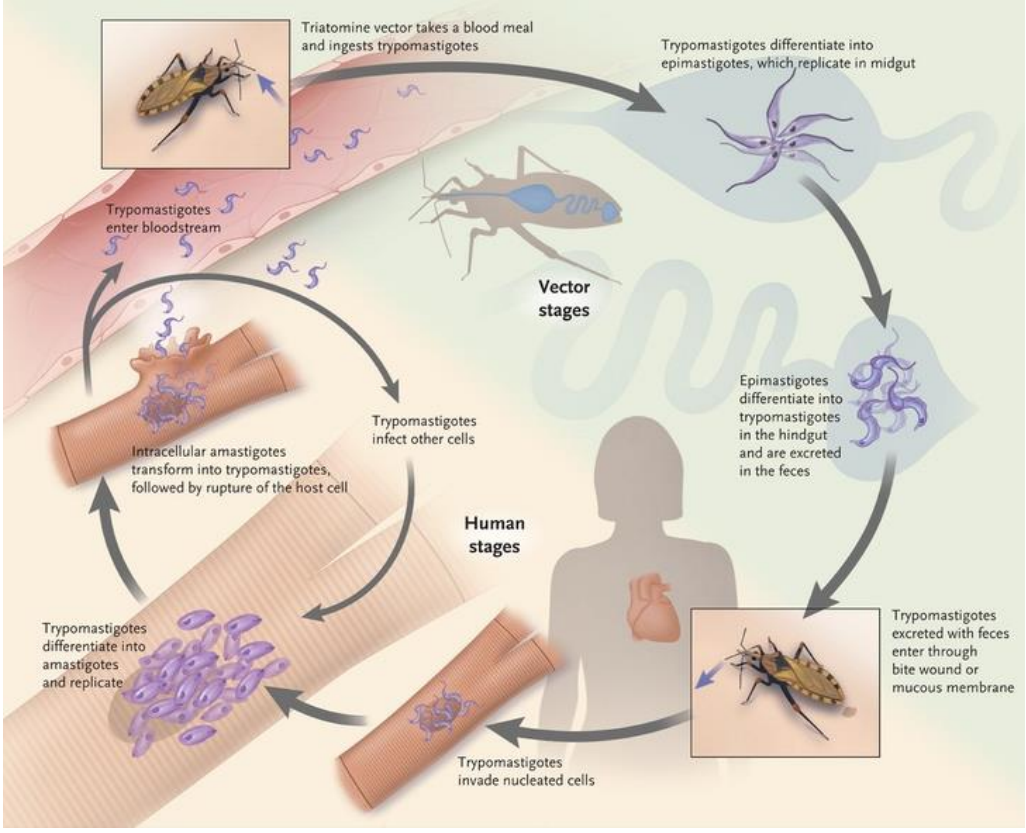

Triatomine vector takes a blood meal and ingests trypomastigotes
Trypomastigotes differentiate into epimastigotes, which replicate in midgut
Trypomastigotes enter bloodstream
Vector stages
Epimastigotes differentiate into trypomastigotes in the hindgut and are excreted in the feces
Trypomastigotes infect other cells
Intracellular amastigotes transform into trypomastigotes, followed by rupture of the host cell
Human stages
Trypomastigotes differentiate into amastigotes and replicate
Trypomastigotes excreted with feces enter through bite wound or mucous membrane
Trypomastigotes invade nucleated cells

Figure 1. Live cycle of Trypanosoma cruzi.

The infestation process occurs through insect feces excretion over mammal skin, promoting the transmission of trypomastigote state. Subsequently, the invasion process takes place in the site of injury (bug bite) and through mucosal membranes. Inside bloodstream, the trypomastigote state colonize different cell types, generally by phagocytosis (Figure 1). Consequently, a series of parasite transformation take place within the colonized cell, arising the non-flagellated amastigote form. Before the cell rupture, the amastigote turns in trypomastigote, and when this latter form arrives to bloodstream it can invade other host cells or can be capture by other vectors. The epimastigote is the replicative form inside the insect, and this form turns in metacyclic trypomastigote during its movement through the insect rectum (Figure 1) (Noireau,Diosque *et al.* 2009). Other forms of parasite infestation could be blood transfusions, organ transplants, through maternal milk, via placental transfer and by the ingestion of contaminated food and drink (Coura,Vinas *et al.* 2014).

2.1.4. Epidemiology of Chagas disease

Chagas disease is endemic of Central and South America, including Mexico. Additionally, enzootic cycles of *T. cruzi* take place in south region of United States (USA) and human infections produced by native vectors have been reported in states of Texas, California, Tennessee, Louisiana and Mississippi (Figure 2) (Coura,Vinas *et al.* 2014;Steverding 2014).

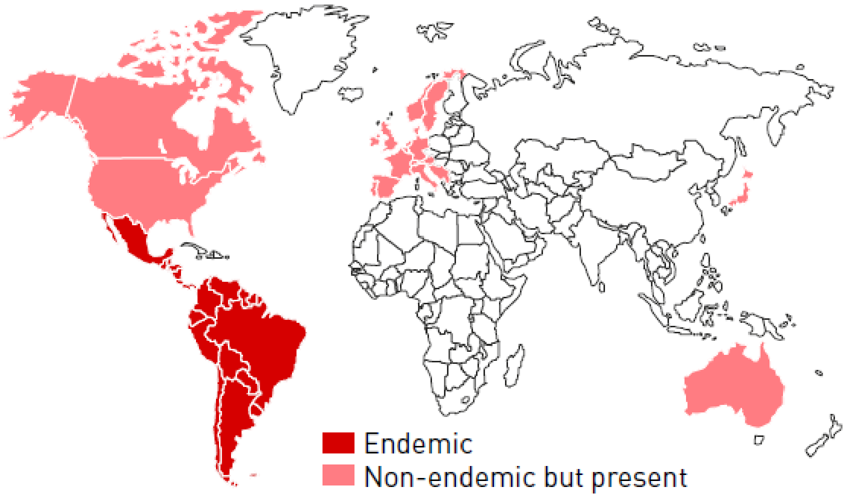


Figure 2. Worldwide distribution of Chagas disease. *The legend shows endemic and emerging areas. (Taken from https://www.dndi.org/diseases-projects/chagas/)*

The epidemiological characteristics and control of Chagas disease vary according to the ecological conditions and health policies of each country. Since the growing of legal and illegal migration from endemic to non-endemic countries, a new socio-politic problem has emerged with the internalization of Chagas disease in regions of North America, Europe, Asia and Oceania. In addition, these migrations have produced new epidemiological problems concerning to public health of countries which receive immigrants. The principal challenging to be handled are the transfusions risks and congenic transmissions. Also, an additional surveillance in blood banks and the specialization of medical staff has been necessary to face the arrival of patients with Chagas in these countries that have few experience on this infection control. Additionally to medical, social and economic problems generated by Chagas disease internalization, outcomes of new politic issue arise in the area of migratory control, because of the immigration stimulation for cheap workforce in develop countries (Coura,Vinas *et al.* 2014;Steverding 2014).

2.1.5. Treatment and control of American trypanosomiasis

Chemotherapy is the main control trypanosomal diseases but *T. cruzi* is not susceptible to many of the drugs used to treat closely related parasites such as *T. brucei*. The treatment for Chagas is based on drugs formulated 50 years ago, which are mainly ineffective, toxic and increase the resistance risk. Although the huge number of annual cases reported with Chagas, a prophylactic treatment does not exist because the disease pathology has an autoimmune component (Wilkinson and Kelly 2009). The chemotherapeutic agents for the treatment of acute phase are nifurtimox (Lampit, Bayer) and benznidazole (Radanil/Rochagan, Roche). However, these drugs are not available in countries such as USA and Canada, excepting those donated by CDC. In addition, these nitroheterocyclics compounds have many side effects and few efficacy for chronic phase treatment (10-20%). Furthermore, several studies report neuropathic and tumorigenic or carcinogenic effects as long-term sequel of these therapeutics drugs. The treatment with the afore mentioned drugs is highly prolonged, needing 2 or 3 pills daily for 60-120 days, which makes that patients discard the treatment in early phase as result of severe side effects (Wilkinson and Kelly 2009).

Currently, another way of this disease control is based on vector surveillance. In order to decrease the biological niches of these insects, a hygienically and environmental strategies have been implemented as well as the use of insecticides in houses to kill the triatomides (Gourbiere,Dorn *et al.* 2012).

There is an urgent necessity to develop new anti-Chagas chemical therapies, in which the patho-pharmacological properties could be improved respect to the available drugs. Because of *T. cruzi* experiments several cycles of intracellular replication, the enzymes that interfere in cell rupture, host invasion and metabolism of this protozoan have become in the principal focus for therapeutic targets. Some intracellular pathways such as polyamine metabolism, sterols and isoprenoids biosynthesis, redox systems, molecules transport, trypanothione synthesis, pentose phosphate pathway and some proteins like trans-sialidase, arginase kinase and proteases in general, constitute a point of growing interest for therapeutic intervention. Currently, several inhibitors have been designed against the molecules which participate in all these biological processes (Wilkinson and Kelly 2009).

2.2. Cysteine proteases

2.2.1. Biochemical aspects of cysteine proteases

Cysteine proteases are those enzymes capable of cleave peptidic bonds through a nucleophilic attack of sulfhydryl group from cysteine residue. The catalytic mechanism is similar to the serine proteases, in which there is necessary the presence of one nucleophile and one proton donor as a base. Like most of serine proteases, one histidine residue has been identified as donor of protons in cysteine proteases (Turk,Stoka *et al.* 2012). These enzymes are widely distributed in nature and came from different evolutionary lineages or clans. The studies about their catalytic mechanism have been performed mainly in papain, which belongs to clan CA. The sulfhydryl group of cysteine and imidazole from histidine generate a catalytic dyad, frequently stabilized by a conserved asparagine. A highly conserved glutamine forms an oxianionic hole, a crucial element for the formation of one stabilizer center of tetrahedral intermediary during hydrolysis (Turk,Stoka *et al.* 2012;Verma,Dixit *et al.* 2016).

The first cysteine protease was purified from *Carica papaya* fruit, which was called papain. According to their structural and functional similarities, the cysteine proteases from diverse origins are divided in three superfamilies. The family of papain, which is the greatest one (also called clan CA), is formed by cysteine proteases of plants (caricain, bromelain and actinidin), mammals (lysosomal proteases such as cathepsin B, L, H, S, C and K) and protozoans (cruzipains, falcipains and *Leishmania* cathepsins (Lmcps)) (Sajid and McKerrow 2002;Verma,Dixit *et al.* 2016).

The structure of papain-like cysteine proteases consists of two domains, one composed predominantly by α-helices and the other by antiparallel β-sheets. The active site, placed at the interface of two domains, encloses the catalytic triad formed by the

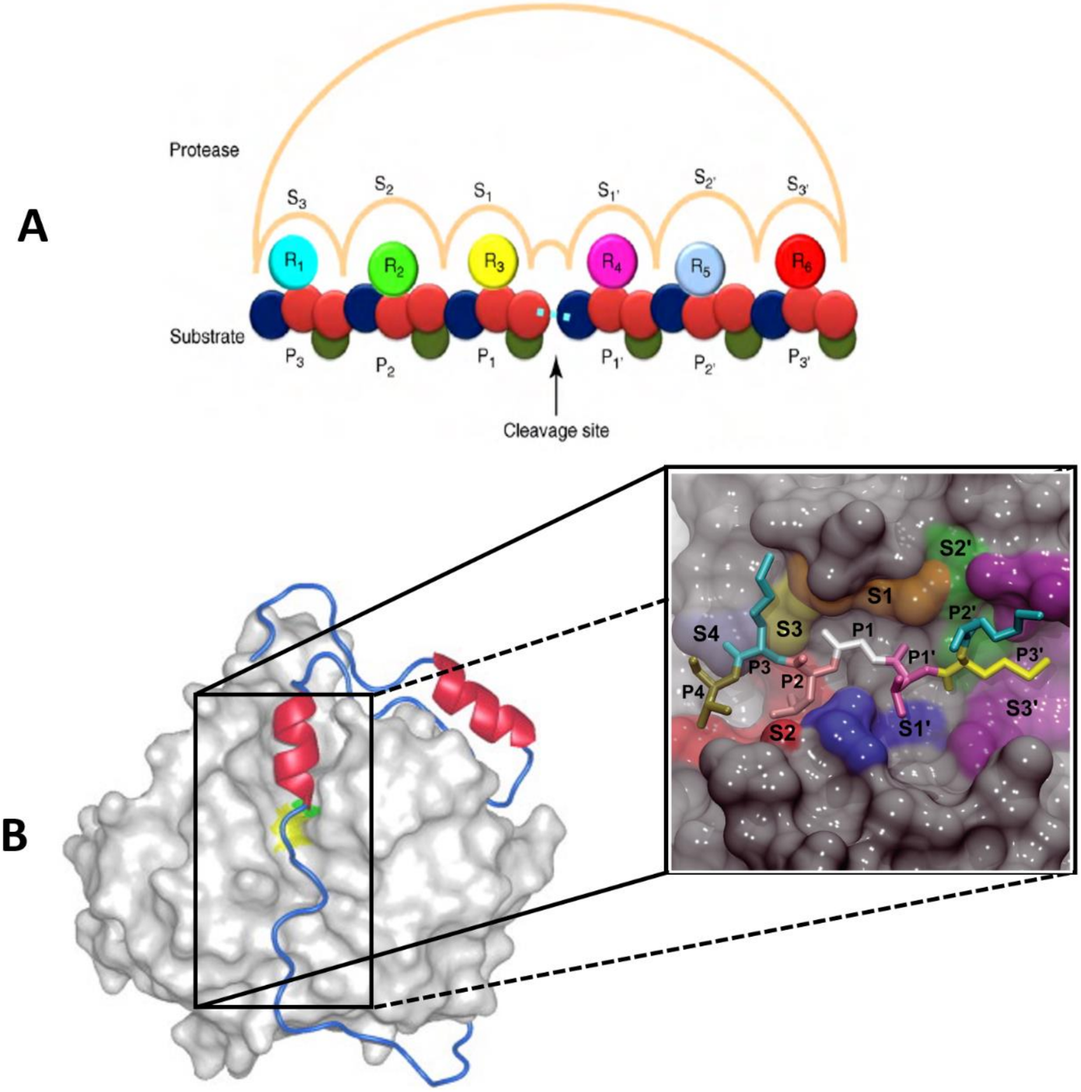


Figure 3. Schematic representation of cysteine proteases binding site. *(A) Scheme of different amino acid residues (P3-P3 ') coupled to the subsites of the enzyme (S3-S3') represented from R1 to R6. (Taken from Smooker et al., 2010 (Smooker,Jayaraj et al. 2010)). (B) Three-dimensional representation of cysteine proteases site in interaction with is propeptidic region. (Taken from Durratn et al., 2010 (Durrant,Keranen et al. 2010))*

residues CYS, HIS and ASN. In past decades (1967), some studies revealed the structure and functionality of papain active site, which is formed by seven subsites (S4-S3’) which interact with polipeptidic substrate residues (P4-P3’) (Figure 3). Since that, one illustrative model was proposed for describing the enzyme-substrate interactions, where each subsite of enzyme (Sx) is defined as the region which interacts with residue (Px) of peptidic substrate. Thus, those residues placed in amino-terminal region starting from cleave site of peptidic bond are named P1, P2,…Pn’, and those residues placed in carboxyl-terminal are named P1’, P2’,…Pn’, respectively (Figure 3) (Drag and Salvesen 2010).

The interactions between polypeptide residues with the residues allocated in S subsites, determine the substrate specificity of the enzyme. However, this simplified model based on key-lock mechanism of pockets or subsites is still inaccurate. This is because the neglecting of electrostatic properties of each subsite which limits the comprehension of protein-ligand interactions. Moreover, the enzyme-ligand complementarity is adjusted by a dynamical process of coupling between flexible chains (Drag and Salvesen 2010).

2.2.2. Cysteine proteases of kinetoplastids as drug targets

The first findings of proteolytic activity in kinetoplastids were reported in decades of 1970 and 1980, discovering a predominant cysteine protease activity. Essentially, the detected activity displayed a behavior similar to observed in human cathepsin L and was localized in endolysosomal system of these parasites. Furthermore, low enzyme levels were detected in the major sites of endocytosis and in the cellular membrane of some kinetoplastids such as *T. congolense* and *Leishmania*. (Mark W. Robinson PhD 2011;Ferreira and Andricopulo 2017).

In the past 30 years, several scientific works have described the functionality and roles of enzymes belonging to CA clan and C1 family within the pathogenesis process and biology of parasitic kinetoplastids. In this sense, it has been demonstrated that expression of proteases with cathepsin L-like activity is regulated throughout different parasite stages in both: intermediary and definitively hosts. Some basic studies of cathepsin genes deletion in *Leishmania* allowed the elucidation of contribution of this enzyme function in processes such as parasite cellular differentiation, virulence and immune modulation. *In vitro* experiments have confirmed that the inhibition of cysteine protease activity by small molecules limit the survival of some kinetoplastids. In addition, critical *in vivo* experiments with *T. brucei* demonstrated that inhibition of cysteine proteases activity with diazomethyl ketone decrease the parasitemia to below detectable levels. All these findings allow that in 1993 the enzymes of C1 family were validated as therapeutic targets of diseases produced by kinetoplastids. Therefore, the progress of inhibitors design against papain-like activity emerges as hoping alternative for the treatment of diseases related to *Trypanosoma* and *Leishmania* genera (Mark W. Robinson PhD 2011;Ferreira and Andricopulo 2017).

Initially, the idea of using parasitic cysteine proteases as therapeutic targets was attractive, but its concretization was obstructed by surrounding fears related to toxicity triggered by inhibition of their C1 homologues in the hosts. However, experimental findings allow the understanding of how to offset these concerns by (*i*) the propensity of some parasites to rapid accumulation of molecules (including drugs) from the environment, (*ii*) the absence of functional redundancy, both qualitative and quantitative, of parasites C1 proteases comparing with those expressed in mammal hosts, and (*iii*) the projection of short-time therapies (one month) which could be sufficient to manage the toxicity. For these reasons, the selective inhibition of cysteine proteases has transcend from laboratory to clinic (Mark W. Robinson PhD 2011).

Synthetic compounds such as vinyl sulfones, non-peptidyl chalcones, acyl hydrazides, amides, thiosemicarbazones and purine-derived nitriles constitute inhibitors of papain-like cysteine proteases with a bioactivity tested in many kinetoplastid parasites. Also, natural products such as diketopiperazines of deep water sediment-derived fungi and triterpenoids have been studied in cathepsins of *T. brucei rhodesiense* (Martinez-Mayorga,Byler *et al.* 2015;Ferreira and Andricopulo 2017).

On the other hand, the study of cathepsins has improved research field of vaccine develop permitting the use of these enzymes as vaccine candidates. In addition, these cysteine proteases constitutes a specific antigen for routine screening diagnostic in several infections produced by kinetoplastids (Mark W. Robinson PhD 2011).

2.2.3. Cruzain as a drug target of Chagas disease

The most characterized cysteine protease of *T. cruzi* has been cruzain, which is known to be essential for parasite survival. Biochemical studies and animal models have validated this protease for control and elimination of *T. cruzi* infection. In this way, 25 crystal structures of cruzain in complex with different ligands have been reported, and the number of scientific articles that report novel inhibitors of this protease have increased since 2000 decade (Sajid,Robertson *et al.* 2011;Ferreira and Andricopulo 2017).

Cruzain is essential to many biological processes and it is expressed in all stages of *T. cruzi* life cycle (Doyle,Zhou *et al.* 2011). This enzyme is found in different cell compartments, such as lysosomes, cell surface, and flagellar pocket, depending on its function and life-cycle stage (da Silva,do Nascimento Pereira *et al.* 2016). In vector stage, cruzain participates in adhesion between *T. cruzi* and the posterior midgut of the insect

(Branquinha,Oliveira *et al.* 2015). Moreover, the overexpression of this enzyme stimulates differentiation from epimastigotes to trypomastigotes. In this stage, the cruzain inhibition cause an accumulation of a toxic protein precursor inside the Golgi complex, which leads to parasite death (Doyle,Zhou *et al.* 2011). Cruzain also contributes to host cell invasion and its inhibition by the endogenous inhibitor chagasin blocks the parasite invasion in muscle cells. Moreover, parasites with low cruzain expression levels are less efficient in infection of laboratory animals (Duschak,Ciaccio *et al.* 2001). Cruzain inhibition also avoid activation of TGF-β signaling cascade in host, which is essential for mammalian cells invasion (Ferrao,d'Avila-Levy *et al.* 2015).

Also, cruzain is a major antigen of *T. cruzi,* which activates several T-cell dependent responses. It has been demonstrated that an increase in titers of anti-cruzain IgG is related to an increase in the disease severity in patients who are in chronic-phase. Cruzain also mediates the proteolysis of immunoglobulins at the hinge region, which leaves Fab fragments covering the surface of the parasite. This process contributes to the escape strategy from the host immune system through the avoid of processes such as opsonization, phagocytosis and activation of the complement cascade (Berasain,Carmona *et al.* 2003). Furthermore, cruzain prevents the activation of macrophages by the degradation of transcription factors, mainly NF-kB, in early infection phase. Altogether, these data confirm that cruzain is an essential component in host-parasite interaction and Chagas disease physiopathology.

As a biochemical singularity, cruzain has been shown to have a broad pH profile ranging from pH 4.5 to pH 9, which is in concordance with its varied cellular and sub-cellular localizations. Also, it has been demonstrated that hydrophobic side chains at P2 positions are preferred at the pH optimum of cruzain (pH 5.5), while the substrate specificity is influenced by pH range near to milieu pH. The importance of GLU205 has been highlighted in substrate specificity in different pH ranges. This because GLU205 is situated in S2 pocket at neutral pH but is directed away at acidic one (Gillmor,Craik *et al.* 1997).

The first X-ray crystal structure of cruzain was solved in complex with a small-molecule inhibitor Z-Phe-Ala-fluoromethylketone. With this atomic structure, the key binding and specificity elements in active site of this enzyme were elucidated and which constituted a starting point for future structure-based drug design (McGrath,Eakin *et al.* 1995). However, other types of inhibitors have been developed based on the structure of compounds which inhibit human homologue proteins. This is the case of inhibitors containing the nitrile functional group, which were deigned based on odanacatib structure, a cathepsin K inhibitor that is soon to be approved for the treatment of osteoporosis (Beaulieu,Isabel *et al.* 2010). Based on these studies, the research team at Merck concluded that neutral molecules are preferable as cysteine protease inhibitors, because they will not concentrate in the acidic lysosomes. Based on this knowledge, two promising compounds, Cz007 and Cz008 (Figure 4), have been developed (Ndao,Beaulieu *et al.* 2014). These compounds are reported as orally active, reversible cruzain inhibitors capable of curing *T. cruzi* infection in an *in vivo* model. In the case of AAE581 (Figure 4), a nitrile-containing peptide, the cruzain subsite that contributes the most to selectivity is the S3 subsite. This compound is more than 1000-times more selective toward cathepsin K *in vivo* over other cathepsins (Turk 2006). Lastly, irreversible α-ketone cruzipain inhibitors, such as the GlaxoSmithK-line compound SB-462795 (Figure 4), which is a heteroatom cyclic ketone, are promising new scaffolds for cruzain inhibition (Turk 2006).

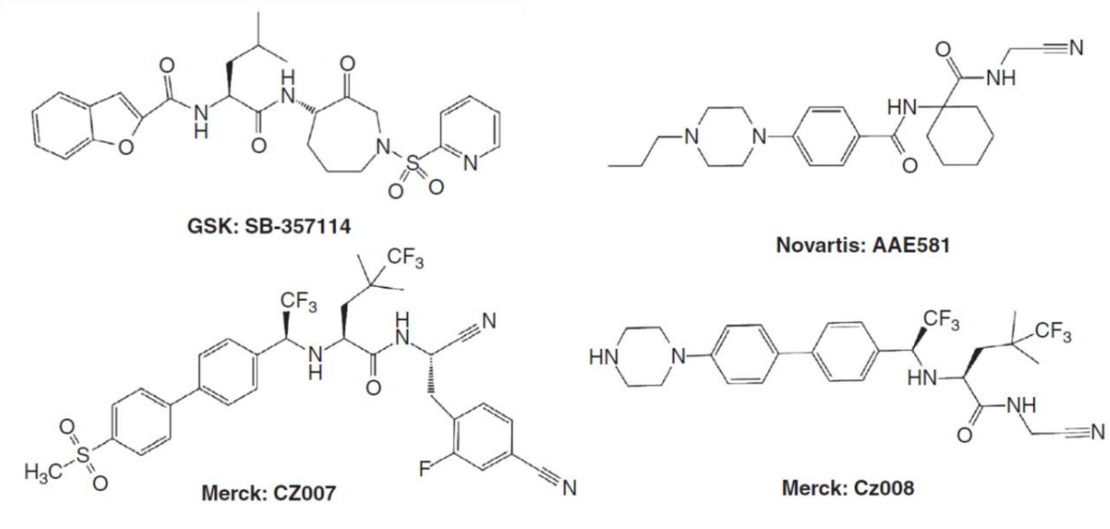


Figure 4. Cruzain inhibitors of potential interest in pharmaceutical industry. *(Adapted from Martinez-Mayorga et al., 2015 (Martinez-Mayorga,Byler et al. 2015))*

The compilation of scientific articles and reported structures comprise about 384 molecules that have been evaluated as inhibitors of this enzyme. Figure 5 shows the most active inhibitors where compounds extensively explored and optimized such as thiosemicarbazones (compounds 2, 3, 12 and 13 in Figure 5) are present. However, the discovery of inhibitors such as isatin (1), oxadiazole (7, 9, 14) and hydrazones (7, 8) has expanded the scaffolds set of chemical structures that can be explored (Figure 5)(Martinez-Mayorga,Byler *et al.* 2015). Recently, the dipeptidyl nitroalkenes have been

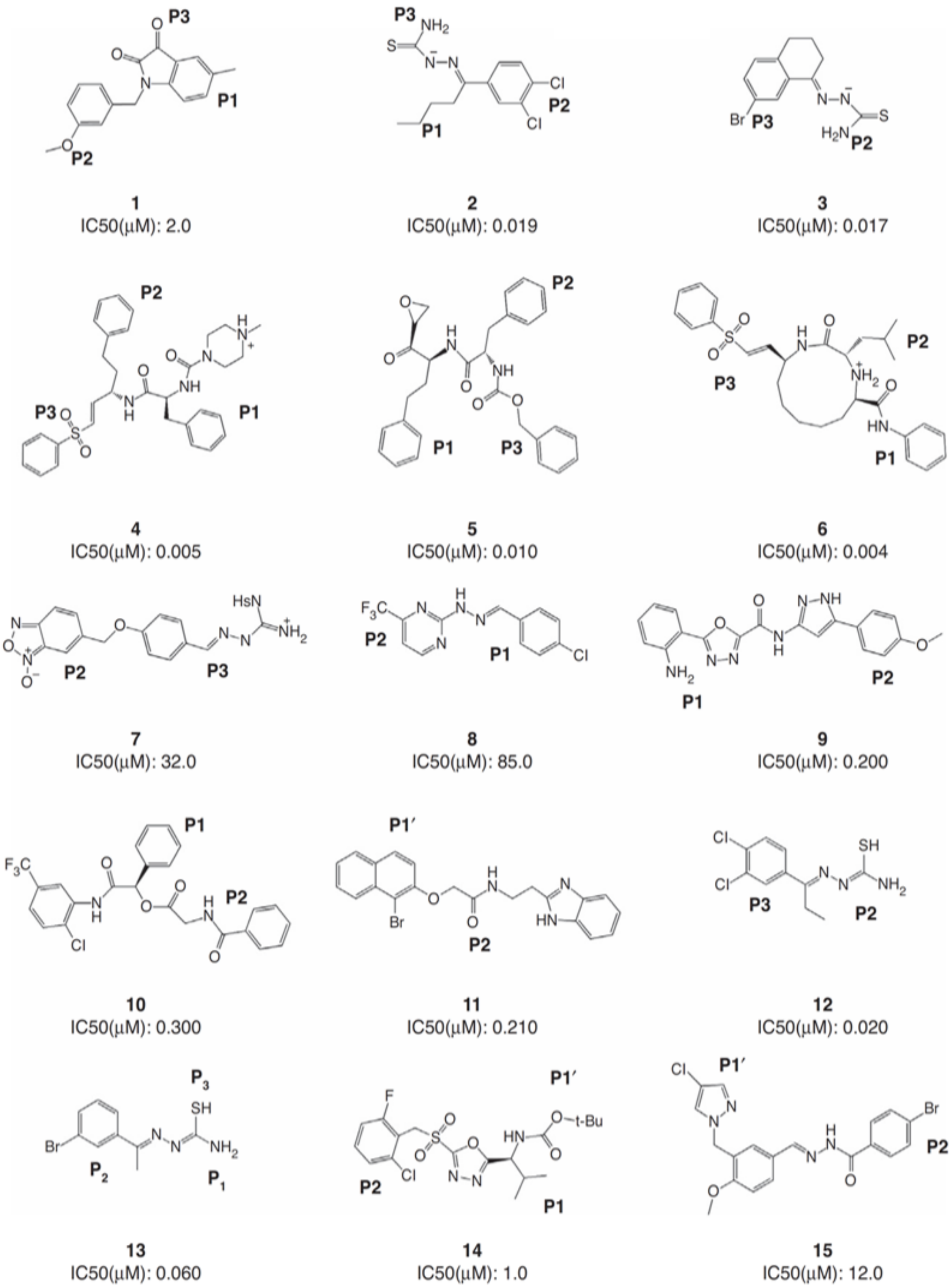


***Figure 5. The most active cruzain inhibitors extracted from several publications.** The chemical groups of compounds are labeled according to their interaction with enzyme subsites. The values of IC50 obtained in the experiments are showed for each case. (Adapted from Martinez-Mayorga et al., 2015 (Martinez-Mayorga,Byler et al. 2015))*

proposed as novel and potent reversible inhibitors of cruzain and rhodesain (Latorre,Schirmeister *et al.* 2016).

Cruzain inhibitors containing vinyl sulphones were developed first by McKerrow *et al*. and provided K777, which showed trypanocidal activity in animal models. This compound inhibits the enzyme by attaching irreversibly to the catalytic CYS25 thiol group (IC50=5 nM) (Doyle,Zhou *et al.* 2007). Although the interruption of the preclinical studies with K777, these pioneering and recent investigations have provided robust evidence for the importance of cruzain as a molecular target in drug discovery of Chagas disease.

2.3. Protein allostery

2.3.1. Allostery: concepts and applications

Allostery is a phenomenon based on the protein interaction with a small molecule through a protein site different from the active one, and this is a basic principle which also controls the activity of many enzymes (Lu,Li *et al.* 2014;Liu and Nussinov 2016) (Figure 6). The first appearance of "allosteric" term was in 1961, when Jacques Monod and Francois Jacob used "allosteric inhibition" to describe a mechanism in which the inhibitor was not a steric analogue of the substrate. Later in the 1960s, two well-known models were proposed to describe allosteric effects: (*i*) concerted MWC model by Monod, Wyman, and Changeux (Monod,Wyman *et al.* 1965) and (*ii*) the sequential KNF model by Koshland, Nemethy, and Filmer (Koshland,Nemethy *et al.* 1966). Since then, conformational change was considered as a signature character in the concept of allostery. However, in 1984, Cooper *et al*. (Cooper and Dryden 1984) described an allosteric model without conformational changes and introduced the term "dynamic allostery," inserting the entropy contribution into the allostery phenomenon. To date, the "conformation ensembles and population shift" and "allosteric networks" have become two major points of view for explaining allostery, which was firstly proposed by Nussinov group in 1999 (Kumar,Ma *et al.* 1999;Ma,Kumar *et al.* 1999;Tsai,Kumar *et al.* 1999).

Three mainstream categories have been developed in order to explain the mechanism of how biological functions are achieved through allostery. The first is based on the principle of thermodynamic equilibrium. The second stablish conceptual models such as conformational selection versus induced fit. Finally, the third embodies numerous implicit and explicit approaches which exploited the inferred structural coupling between the functional (active) and allosteric binding sites in a host protein. These implied that structural linkage is a necessary condition for an allosteric action. However, the structural view of allostery has been questioned since according to one of the thermodynamic models detailed structural information is not required (Hilser,Wrabl *et al.* 2012).

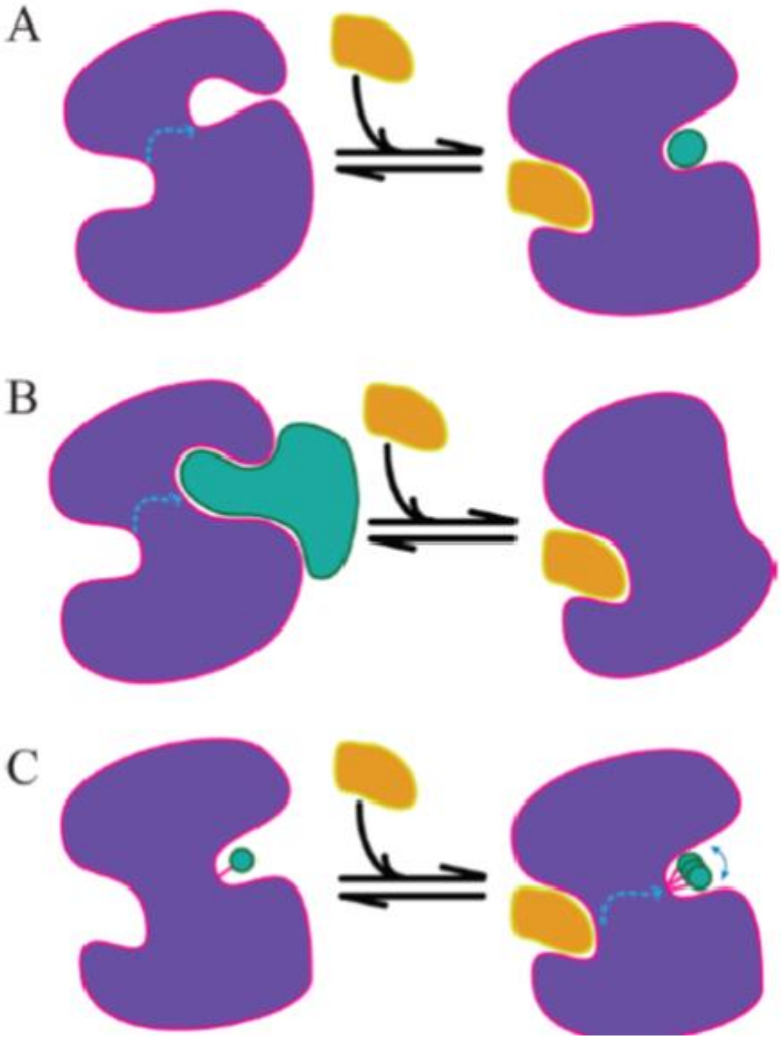


Figure 6. Several forms of allostery regulation in proteins. *(A) Allosteric modulator drives the ligand binding: the association of one ligand (yellow) induces conformational change in a protein (purple) that makes it competent to bind another ligand (green). (B) Allosteric modulator drives disruption of protein−protein interactions: the association of a ligand induces conformational change in a protein that eliminates interaction with a partner protein. (C) Allosteric modulator drives dynamic coupling: the association of a ligand induces changes in the dynamics of the active site (green) thereby affecting its activity. (Taken from Dokholyan et al., 2016 (Dokholyan 2016))*

All these aforementioned categories can be clarified by the basic physical principle that biomacromolecules consist of ensembles of conformations with a certain distribution, which can be described by their free-energy landscape (Figure 7) (Tsai and Nussinov 2014). However, this distribution is not static and could be altered by some environmental changes, such as covalent or noncovalent binding, pH, temperature, or ionic force. Proteins may adopt several conformations of which many are active and others are inactive. There is a balance between these states and they can be interconverted after a reversible interaction of a ligand or irreversible binding to the allosteric site, resulting in a redistribution of the set of protein conformations (Figure 7) (Tsai and

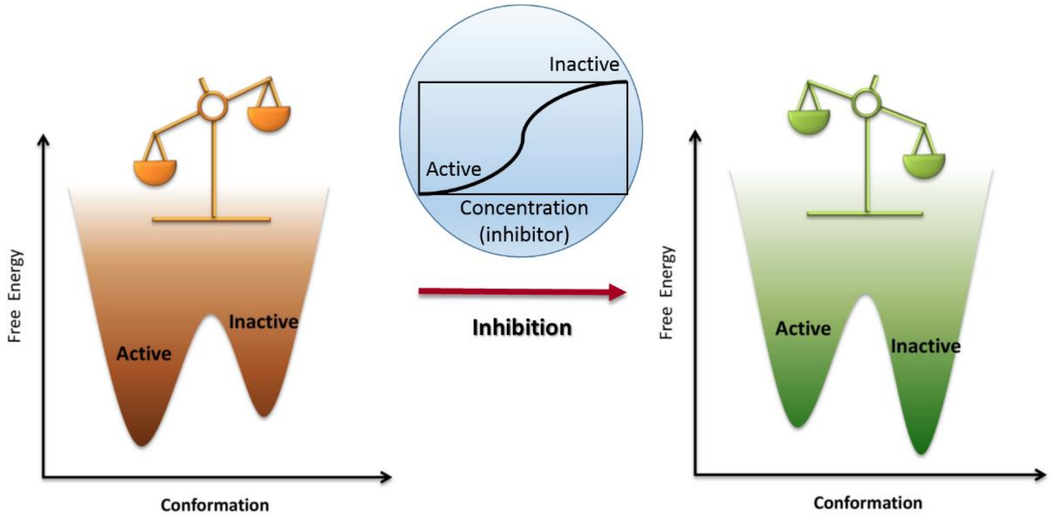


Figure 7. A typical allosteric inhibition via a bi-stable switch. *A protein is illustrated by only two populated states, active and inactive, separated by a sizeable free energy barrier. Before inhibition, the active state dominates the population as indicated by the relative bowl depth in the free energy landscape and a balance level. Within a narrow increment range in ligand concentration, the allosteric inhibition event shifts the population in favor of the inactive state. The inhibition is highlighted in the embedded plot with a typical sigmoid transition from the active to the inactive state. (Adapted from Tsai et al., 2014 (Tsai and Nussinov 2014))*

Nussinov 2014;Nussinov 2016). This “population shift”, occurs principally by the relative stability of conformation changes, and the distribution of these populations on the energy landscape are determined by statistical thermodynamics. The free energy landscape representation effectively describes active and inactive states and their relative stabilities, which helps to understand protein behavior under physiological conditions and dysfunction of mutants in disease (Figure 7).

Nevertheless, proteins are not the only biological molecules that feature allosteric transitions. It has been demonstrated that allostery phenomenon is an inherent characteristic of most macromolecules (Tsai and Nussinov 2014;Liu and Nussinov 2016;Nussinov 2016). On the other hand, allostery is not the sole means for conformational selectivity in the cell. Concentration, availability, micro-organization etc. are all fundamental parts of the cellular machinery. The novel theory about allostery, which is based on the energy landscape, inspired new experimental approaches for characterizing functionally relevant ensembles in proteins. Therefore, the scientific community has been challenged to understand the mechanisms in biology with a point of view differing to the simple interpretation of the structure−function paradigm (Tsai and Nussinov 2014).

One of most popular applications of allostery is to understand the mechanism of biological systems. Therefore, in the past decades, this phenomenon has been deeply studied by experimental and computational methods (Nussinov and Tsai 2013;Nussinov and Tsai 2015). X-ray crystallography is one of the most frequently used experimental methods. It can provide detailed protein structural analysis in the absence or presence of allosteric effectors allows the characterization of the modulation site and establishes the molecular basis of the active site rearrangement (Nussinov and Tsai 2015). However, allostery is a dynamical process and the lack of dynamical information on the static crystal structure limits on X-ray crystallography in studies of this phenomenon. On the other hand, nuclear magnetic resonance (NMR) spectroscopy capture more “snapshots” on transient conformations that are less populated. Computational approaches also complement experimental methods and provide powerful tools to study allostery. The large number of snapshots generated from molecular dynamics simulations captures the motion of the proteins, thus providing insights into the population shift of the protein conformational ensemble (Nussinov 2016).

2.3.2. Allostery in drug development

Allosteric inhibitors are those that decrease the protein’s activity, producing dynamic changes and the losses of protein function upon the binding of a molecule at the protein’s allosteric site. This kind of regulation has produced great interest in the medicinal chemistry area to improve compound selectivity, making the study of allosteric modulation an emergent strategy in the field of drug discovery. The design of allosteric drugs has emerged as a promising research field in disease control because they offer a noninvasive protein regulation. Allosteric regulators do not interfere with endogenous regulators that bind to binding sites or its proximity, they can function in concert with direct active site regulators, and can serve as activating modules. In addition, allosteric effectors may have spatiotemporal specificity (Nussinov and Tsai 2013;Lu,Li *et al.* 2014;Dokholyan 2016;Wagner,Lee *et al.* 2016). For example, they can be active only in presence of an endogenous target, thus restricting their effect to certain tissues at certain times (Groebe 2006;Groebe 2009;Kenakin 2010).

Since allosteric cavities are typically less evolutionarily conserved, allosteric drugs can be highly selective, even among other members of the same protein family (Nussinov and Tsai 2012;Grover 2013;Ma and Nussinov 2014). Interestingly, in some cases, allosteric pockets are so unique among proteins that an effector is said to have "absolute specificity" (Gunasekaran,Ma *et al.* 2004;Groebe 2006;May,Leach *et al.* 2007;Wenthur,Gentry *et al.* 2014). Hence, allosteric sites present a unique and focused opportunity to selectively target one particular protein in a family of homologous proteins. In addition, the ability to control a single protein's function among homologous or functionally duplicating proteins presents an opportunity to scan into cellular connections at the single protein level. Moreover, such control should not disrupt the native function of the target protein, thus making allosteric control a unique handle on protein function.

Similar to orthosteric drugs, allosteric drugs can be classified as (*i*) noncovalent or (*ii*) covalent. A regime of low-dosage of allosteric drug using noncovalent binders are likely to be effective if the protein displays a ''conformational switching'' mechanism between the active and inactive conformations. Drug binding could lead to a shift in the free-energy landscape toward the inactive conformation (Figure 7). Later, in the absence of the drug, if the energy barriers between the states are high, switching back to the active conformation may require long timescales, which may be beneficial in leading to lower drug dosages. Covalent allosteric drugs are, by contrast, more likely to display irreversible action even though this reversible/irreversible distinction is not absolute (Nussinov and Tsai 2013).

Despite many potential advantages of allosteric therapeutics, to date the development of novel approaches for discovering allosteric drugs remains to be a challenge. In recent decades, the pharmaceutical industry has favored more traditional targets for several reasons: (*i*) the relative ease of assay development around orthosteric sites, (*ii*) access to high-throughput, and (*iii*) advances in ligand- and receptor-based computational methods to optimize ligand-binding affinity at a substrate-competitive site. Also, the structure-based approach is thought to significantly reduce the time and cost of hit-to-lead and lead-to-drug development by reducing the number of compounds that need be synthesized (Muchmore and Hajduk 2003;Wagner,Lee *et al.* 2016).

The identification of allosteric sites in proteins that constitute therapeutic targets is the first step in the discovery of drugs that mimic this type of regulation. However, the experimental methods used to characterize these supramolecular structures have faced major challenges such as the increase in the number of therapeutic targets and the limited number of compounds that may or may not detect allosteric sites. In this way, several *in*

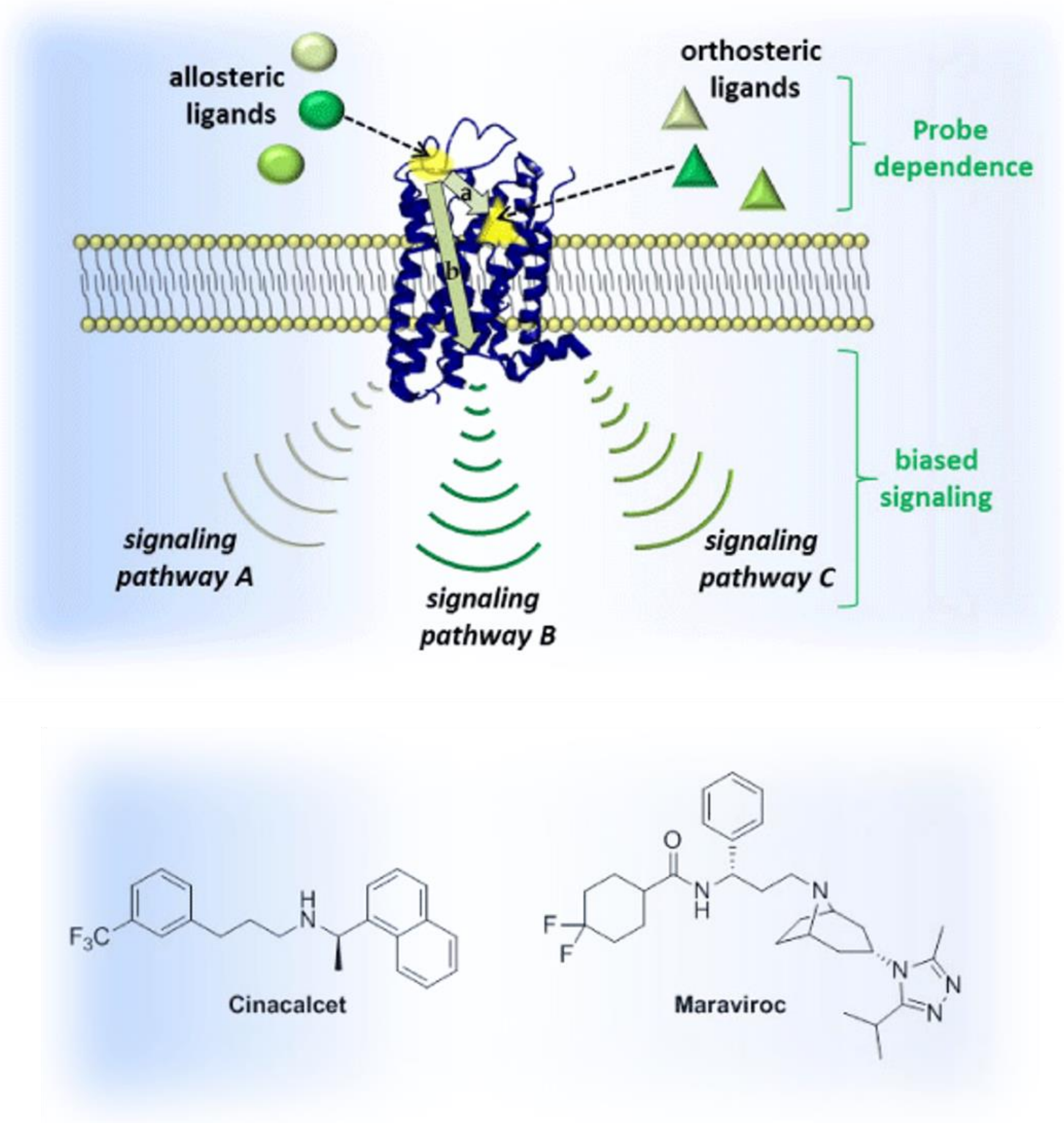


Figure 8. Schematic of GPCR signal transduction and the approved allosteric drugs against these receptors*. Allosteric ligands (spheres; shading represents structural diversity of different allosteric modulators that can act at same site) bind to a site (yellow sphere) that is distinctly different than the orthosteric site (yellow triangle) which accommodates an orthosteric ligand (triangles; shading represents structural diversity of different ligands that can act at the same site). The chemical structures represent the FDA–approved GPCR allosteric modulators. (Taken from Wild et al., 2104 (Wild,Cunningham et al. 2014))*

silico strategies have been developed as an alternative to identify allosteric sites. These methodologies can employ the protein sequence analysis and the study of the dynamic properties of protein crystal structures (Panjkovich and Daura 2012;Collier and Ortiz 2013;Lu,Huang *et al.* 2014;Stank,Kokh *et al.* 2016).

In contrast, allosteric drugs are challenging from a rational drug-design perspective. Because experimental assays typically measure orthosteric function rather than ligand binding at the allosteric site, the efficient development of allosteric drugs requires that the complex structure−activity relationships governing both binding affinity and allosteric activity be considered simultaneously. While the less evolutionary conservation of allosteric sites increase subtype specificity, the chances of evolved resistance are more elevated (Nussinov and Tsai 2012;Wenthur,Gentry *et al.* 2014;Wagner,Lee *et al.* 2016) and can complicate testing in evolutionarily distant animal models (Wenthur,Gentry *et al.* 2014). In addition, optimizing allosteric modes of action requires methods that are very different from those used in orthosteric drug discovery (Nussinov and Tsai 2012). While drug designers may desire to target a single protein function, an allosteric effector may also alter other functions, hindering a full mechanistic understanding of the pharmacology (Wenthur,Gentry *et al.* 2014). Lastly, the assessment of limited number of known allosteric pockets indicates that they are generally shallow (Nussinov and Tsai 2012).

Notwithstanding all these challenges, allosteric drug discovery has gained momentum recently due to a number of developments (Wenthur,Gentry *et al.* 2014). First, several allosteric drugs across a broad range of pharmacological target classes have been rationally designed, (Rawal,Murugesan *et al.* 2012;Saalau-Bethell,Woodhead *et al.* 2012;Gentry,Sexton *et al.* 2015;Lugowska,Kosela-Paterczyk *et al.* 2015;Wu,Nielsen *et al.* 2015;Meng,McClendon *et al.* 2016) encouraging search of others, as evidenced by the number of allosteric drugs currently in clinical trials (Wu,Nielsen *et al.* 2015). Finally, advances in our understanding of allosteric mechanisms have supported the development of additional rational design strategies. An illustration of success in the field of rational drug design of allosteric modulators is the classic example of allosteric drugs approved by the FDA (Food and Drug Administration agency) against GPCR (G Protein-Coupled Receptors) (Figure 8) (Wild,Cunningham *et al.* 2014).

2.3.3. Allosteric regulation in papain-like cysteine proteases

In general, there is growth trend in the number of publications which highlight the importance of allosteric modulation in proteases. However, most of reports suggest that in this kind of enzymes, allostery is mostly mediated by protein-protein interactions during the assembly process. This makes the small molecule-mediated allostery an incipient subject within the field of proteases enzymatic control (Hauske,Ottmann *et al.* 2008;Shen 2010).

Several evidences have demonstrated the allosteric regulation of cysteine proteases, in mammals as well as in some parasites enzymes (Marques,Esser *et al.* 2013;Novinec,Korenc *et al.* 2014;Marques,Gomes *et al.* 2015). Interestingly, cysteine peptidase cathepsin K (HuCatK) is becoming established as a model enzyme for regulation of papain-like peptidases via sites distant from the active site (Novinec,Korenc *et al.* 2014;Novinec,Lenarcic *et al.* 2014;Novinec,Rebernik *et al.* 2016;Novinec 2017). However, lack of significant conformational change in X-ray structures has thus far hindered our understanding of this system (Novinec,Rebernik *et al.* 2016;Novinec 2017). Also, the identification of sectors that mediate allosteric communication within this protease family was performed by SCA (Novinec,Korenc *et al.* 2014). Therefore, all allosteric sites were predicted by mapping the molecular surface components directly in contact with these sectors. These results allowed the designing of two allosteric inhibitors of HCatK, and the crystal structures of both complexes were also obtained (Novinec,Korenc *et al.* 2014;Novinec,Rebernik *et al.* 2016).

In addition, several crystallographic structures that characterize the interaction of human cathepsin K with chondroitin sulfate (PDB: 3C9E, 3H7D and 4N8W) have been solved. Interestingly, numerous studies have also reported the allosteric modulation of various papain-like cysteine proteases by GAGs (Almeida,Nantes *et al.* 1999;Almeida,Nantes *et al.* 2001;Lima,Almeida *et al.* 2002;Li,Yasuda *et al.* 2004;Novinec,Kovacic *et al.* 2010;Costa,dos Reis *et al.* 2012;Judice,Manfredi *et al.* 2013). Therefore, the prediction of GAGs binding sites in parasitic cysteine proteases could be the starting point for its characterization and therapeutic intervention. In addition, the enzymatic activity inhibition of falcipain 2 and 3 from *Plasmodium falciparum* by the heme and suramin was also demonstrated, suggesting that this inhibition occurs through an allosteric binding mechanism (Marques,Esser *et al.* 2013;Marques,Gomes *et al.* 2015).

3. Methodology

3.1. Theoretical foundations

3.1.1. Molecular docking and virtual screening

The molecular docking consists in the prediction of binding mode of a certain molecule into the binding site of another target molecule. Docking algorithms can cope with different of biomolecular complexes, such as protein-protein, protein-DNA or protein-ligand. Figure 9 depicts the prediction of a protein-ligand complex structure. Most docking algorithms involve two steps: (*i*) the generation of different conformations of the binding partners in the bound state, and (*ii*) the ranking of these conformations through energy scoring functions (Yuriev,Holien *et al.* 2015).

Molecular docking is often applied in the drug design field. It allows the rational selection of compounds with potential pharmacological activity, based on the predictions of binding modes and affinities of the ligands and/or substrates to the target protein. Therefore, molecular docking is a useful tool for the virtual selection of inhibitors and may reduce the costs and time required for the hit identification (Lionta,Spyrou *et al.* 2014;Ferreira,Dos Santos *et al.* 2015).

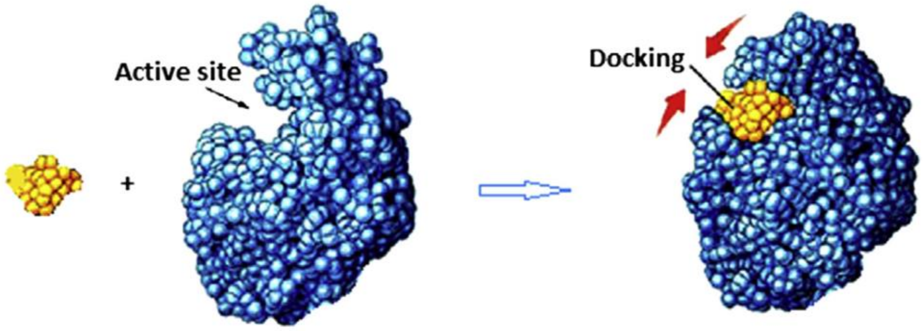


Figure 9. Molecular docking of drug (yellow) into the binding site of the receptor (blue).

Most docking algorithms typically comprise two recognition models (*i*) lock and key, and (*ii*) induced fit. The interactions are modeled through force-fields which take into account van der Waals and electrostatic terms, as well as the formation of hydrogen bonds. Some programs may also include entropic and solvation effects (Yuriev,Holien *et al.* 2015).

The docking precision is determined by various factors, i.e., the limited resolution of the target molecules, the flexibility of the binding partners, the occurrence of induced-fit or other conformational changes during the binding process, and the presence of water-mediated interactions at the interface. Most docking protocols take into account the flexibility of the ligand while keeping the protein conformation fixed, which is known as rigid docking, in opposition to flexible docking, where some user-defined interface residues are able to move (Chen 2015;Yuriev,Holien *et al.* 2015).

Docking algorithms are complemented by scoring functions that estimate the binding quality of ligand. They are based on energetic functions, which include electrostatic and van der Waals terms. Subsequently, there is classification of conformations (ranking) where they are revalued and the binding free energy is estimated, in which in some programs the entropic and solvation effects are included (Gohlke and Klebe 2002).

The treatment of ligand flexibility is divided into the following categories: systematic methods; stochastic methods or deterministic methods (Brooijmans and Kuntz 2003). The systematic search attempts to systematically explore all degrees of freedom in a molecule. In order to avoid exhaustive search calculations, many algorithms use an increased ligand construction (Rarey,Kramer *et al.* 1996), which consists of dividing the ligand into a rigid region and flexible lateral chains defined by the identification of bonds with free rotation. The rigid part is first anchored and the flexible parts are sequentially added, with a systematic scan of the torsion angles. Another method of systematic search is the use of pre-generated conformation libraries. In this way, the process becomes in a docking of a "rigid body". Stochastic methods make random changes in a ligand structure or a ligand population. The resulting structure is evaluated according to a probability function. The most common methods are Monte Carlo and genetic algorithms.

Most docking algorithms are optimized to be fast, which allows the screening of large compound databases. On the other hand, the development of new docking tools has led to high quality programs. Many of them can be found on the website click2drug.com, which contain a summary of useful molecular modeling tools related to rational drug design (Ou-Yang,Lu *et al.* 2012).

It is worth noting the high costs associated to the reagents and laboratory material employed in high-throughput screening of large commercial compound libraries. Therefore, instead of performing experimental assays in the initial stages of the drug discovery projects, a high throughput screening of compound libraries can be carried out in order to select *in silico* a handful number of potential hits that may undergo subsequent experimental validation steps. This *in silico* step is called virtual screening (VS), in which thousands or millions of compounds are docked into a crystallographic structure of a macromolecule. The VS results consist of ranking list where the compounds having the highest scores/affinities for the target molecule occupy the first positions. Even though some results of VS are not confirmed during experimental validation, there are in the literature several successful applications of this methodology in drug design (Schneider 2010). On the other hand, despite the successful applications, various groups are currently working to improve the performance of VS procedures. In this sense, the ensemble-based approach is one of the best current ways to cope this technique considering the receptor flexibility in (Korb,Olsson *et al.* 2012).

3.1.2. Classic molecular dynamics

The tridimensional structures of biomolecules are mostly obtained through experimental techniques such as NMR and X-ray crystallography, or through homology modeling. These techniques generate a static representation of the molecular structure. Though useful, the information gathered by the previous approaches cannot deal with the molecular movements, which are essential to the molecules' functions. In this context, the molecular dynamics (MD) simulations represent a powerful computational tool to study the properties of biomolecules in equilibrium, their movements and transport. MD simulations can provide a quantitative description of time-dependent fluctuations, conformational changes in proteins and other functions in multiple biological systems (Purdy,Bennett *et al.* 2014;van den Bedem and Fraser 2015). On the other hand, the essence of structure-based drug design is the comprehension of protein-ligand interactions. Therefore, the MD simulations have been extensively applied in this specific field. Numerous studies have described the usefulness of MD simulations to understand the binding modes of known drugs, in order to guide the design of new inhibitors.

MD simulations are based on the classical mechanics, which considers the atoms as solid particles connected by springs (covalent bonds) (De Vivo,Masetti *et al.* 2016). The springs oscillate around an optimal or equilibrium distance, i.e., the average bond length. The particles obey the Newton's equation of motion:

$$F_i = m_i \frac{\delta^2 ri}{\delta t^2} \qquad (for\ i = 1, 2, \dots N) \tag{1}$$

where F_i, m_i y r_i are the force acting on the *i*th particle, its mass and position, respectively, and *t* stands for the time. The force acting on the *i*th particle can be calculated by the derivative of the potential energy function *U* (*r1, r2,…, rN*) as follows:

$$F_i = -\frac{\delta U}{\delta r_i} \tag{2}$$

The acceleration is calculated employing the forces and the masses:

$$a_i = -\frac{F_i}{m_i} \tag{3}$$

An example of an integrator, the velocity-Verlet is a simple and one of the first used algorithm in MD simulations (Verlet 1967). Within such an integrator, positions at time (*t-δt*) are calculated by current positions, velocities (*v(t)*), and acceleration according to following:

$$x_i(t + \delta t) = x_i(t) - v_i\,(t)\delta t + \frac{1}{2}a_i\,(t)\delta t^2 \tag{4}$$

and velocities are propagated as follows:

$$v_i(t + \delta t) = v_i(t) + \frac{1}{2}[a_i(t) + a_i(t + \delta t)]\delta t \tag{5}$$

An MD simulation is usually carried out through an iterative process as follows: (*i*) definition of the initial coordinates and velocities of all the atoms in the system, (*ii*) computation of forces acting on all the atoms of the system using the force-field equations, (*iii*) integration of Newton's equation of motion. After every discrete step, the coordinates, forces and velocities are updated and a new iteration starts at *t*+δ*t.* The previous process occurs until the system reaches the user-defined simulation time. One of the most important requirements for the iterative MD simulations is an efficient method to solve the equation of motion. The integration is carried out iteratively at predefined δ*t* step. This value cannot be smaller than the faster movements of the systems, i.e., the vibration of H atoms around their equilibrium bond lengths. Therefore, to increase the δ*t* in order to obtain larger simulation times, one typically apply constraints on the covalent bonds involving H atoms with methods like LINCS or SHAKE. By doing so, one increases the δ*t* up to 2 fs (De Vivo,Masetti *et al.* 2016).

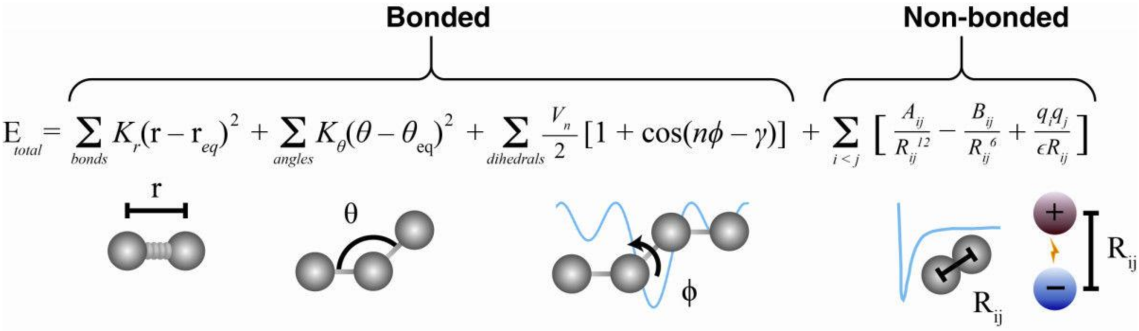

Bonded
Non-bonded
$E_{total} = \sum_{bonds} K_r(\mathrm{r} - \mathrm{r}_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} [\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}]$
r
θ
ϕ
R_{ij}
+
−
R_{ij}

Figure 10. Bonded and non-bonded interactions included in classical force fields. *The atomic forces that govern molecular movement can be divided into those caused by interactions between atoms that are covalently bonded and those which are not bonded. Chemical bonds and atomic angles are modeled using simple springs, and dihedral angles are modeled using a sinusoidal function that approximates the energy differences between eclipsed and staggered conformations. Non-bonded forces arise due to van der Waals interactions, modeled using the Lennard-Jones potential, and electrostatic interactions, modeled using Coulomb's law. (Taken from Durrant et al., 2011)*

The force experienced by each atom in the system during and MD simulation is given by the force-field equations, which takes into account the interactions of each atom with the rest of the system (Figure 10). Force fields are mathematical expressions that contain the functional forms of the potential energy which, in turn, includes bonded and non-bonded interactions among the atoms of the system (Figure 10). Such interactions are described by terms related to: (*i*) covalent bond stretching, (*ii*) angular bending, (*iii*) dihedral torsion, (*iv*) van der Waals interactions and (*v*) electrostatic interactions.

The bonded terms (bonds and angles) are modeled employing harmonic potentials, the dihedrals are modeled by employing periodic (sine/cosine) functions and the non-bonded terms (van der Waals and Coulomb) are modeled through Lennard-Jones and Coulomb potentials, respectively. A distance cutoff is necessary to defined which atoms will interact with a given atom *ith* through non-bonded interactions.

3.1.3. Binding free energy calculations in protein-ligand systems

Methods to calculate free energies of binding from molecular simulation have been around for more than two decades. The advantages of calculating protein–ligand interactions in a thermodynamically way were recognized immediately. Accurate simulations offer insight at a molecular level, which are valuable for rationalizations and predictions in drug design processes (de Ruiter and Oostenbrink 2011). With ever increasing computational power and methodological advances, these methods are being applied more and more often. The main advantage over faster, empirical scoring functions being a correct inclusion of all thermodynamically relevant phenomena, like protein flexibility and explicit inclusion of water. Moreover, free energy methods include both energetic and entropic contributions (de Ruiter and Oostenbrink 2011).

Several computational methods, from computationally rigorous thermodynamics pathways approaches to less complex end-point methods, have been developed for ΔG_{bind} calculations. The former methods include thermodynamic integration (TI) and free energy perturbation (FEP) methods, whereas liner interaction energy (LIE), MM-generalized Born surface area (MM-GBSA), and MM-Poisson–Boltzmann surface area (MM-PBSA) are end-point methods. Each of these methods has its own strengths and limitations, and their computational requirements and speed are inversely correlated with their accuracy. Understandably, TI and FEP methods demand multiple MD simulations and rigorous sampling of ligand, protein, and solvent degrees of freedom. As a result, the thermodynamic pathways methods are in general able to provide accurate estimates of the free energy of binding at a cost of high computational time (Wereszczynski and McCammon 2012;Rathore,Sumakanth *et al.* 2013;Christ and Fox 2014).

Less rigorous alternatives to thermodynamic pathways are the end-point approaches, which sample only structures involved at either ends of the reaction pathways; that is, the free receptors (proteins) and ligands and the final ligand–protein complexes.

MM-GBSA and MM-PBSA are computationally efficient methods for estimating the binding free energy (ΔG_{bind}) of protein-ligand complexes (Bashford and Case 2000;Hou,Wang *et al.* 2011;Homeyer and Gohlke 2012;Kleinjung and Fraternali 2014). This thermodynamic state function is calculated as shown in equation 6.

$$\Delta G_{bind} = G_{complex} - \left[G_{protein} + G_{ligand} \right] \tag{6}$$

Here, ΔG_{bind} correspond to the free energy difference between the bound and unbound states of a complex. The former terms are calculated as follows in the MM-GBSA formalism:

$$G = \langle E_{MM} \rangle + \langle G_{sol(polar+nonpolar)} \rangle - T\langle S_{MM} \rangle \tag{7}$$

where E_{MM} is the Molecular Mechanical energy of the molecules in gas phase. E_{MM} is the sum of the internal energy (E_{int}) of the molecules (i.e. bonded terms), E_{elec}, and

E_{vdW}, which represent the intermolecular electrostatic and van der Waals interactions, respectively.

$$E_{MM} = E_{\text{int}} + E_{vdW} + E_{elec} \tag{8}$$

G_{solv} is the sum of the polar (G_{GB}) and non-polar solvation (G_{SA}) energies (equation 9). The G_{GB} contribution is calculated by solving the Generalized Born (*GB*) equation of the molecules and G_{SA} through the solvent accessible surface area (SASA) method for the non-polar part of the solvation energy. T is the temperature and S_{MM} is the entropy.

$$G_{sol} = G_{polar} + G_{nonpolar} = G_{GB} + G_{SA} \tag{9}$$

In the Poisson-Boltzmann model (*PB*), the polar contribution is computed through the well-known *PB* equation. On the other hand, in the case of *GB* models, the polar solvation component (G_{GB}) is calculated through equation 10 proposed by Still *et al* (Noireau,Diosque *et al.* 2009). Even though the *PB* model is considered as a more rigorous approach, *GB* models are less computationally-demanding and often give fairly satisfactory predictions (Hou,Wang *et al.* 2011;Zeller and Zacharias 2014).

$$G_{GB}\left(X\right) = \frac{1}{2}\left(1\frac{1}{\varepsilon_{\omega}}\right)\sum_{i,j} q_i q_i \left[r_{ij}^2 + R_i R_j \exp\left(\frac{r_{ij}^2}{4R_i R_j}\right)\right]^{\frac{1}{2}} \tag{10}$$

The term εw is the dielectric constant of the solvent (e.g. water). *i* and *j* represent the solute atoms, being r_{ij} the distance between them, q_i and q_j, their partial charges, and R_i and R_j, their effective Born radii.

The non-polar solvation contribution is calculated through equation 11, where *SA* stands for the solvent-accessible surface area of the solute. Coefficients γ and β are empirical constants with values of 0.0072 kcal/mol and 0, respectively, for the *GB* models (Jayaram,Sprous *et al.* 1998).

$$G_{SA} = \gamma SA + \beta \tag{11}$$

Energetically-relevant residues, i.e., warm- and hot-spots, at the interfaces of the studied complexes were predicted by using the per-residue effective free energy decomposition (*pr*EFED) protocol implemented in MMPBSA.py, a python program which is within the repository of Ambertools package (Miller,McGee *et al.* 2012). Of note, we defined as warm- and hot-spot residues those with a side-chain energy contribution (ΔG_{sc}) to the total ΔG_{eff} value ranging from -1.0 to -0.4 kcal/mol and ≤-1.0 kcal/mol, respectively, as defined elsewhere (Humphris and Kortemme 2008). The per-residue free energy contribution (ΔG_{res}) under the single trajectory approach is calculated as follows (Gohlke,Kiel *et al.* 2003;Zoete and Michielin 2007):

$$\Delta G_{res} = \frac{1}{2}\sum_{i\in res,\, j\notin res}\left(\Delta E_{vw}^{ij} + \Delta E_{el}^{ij} + \Delta G_{GB/PB}^{ij}\right) + \sum_{i\in res}\Delta G_{GB/PB}^{ii} + \Delta G_{SA(res)} - T\Delta S_{res} \quad (12)$$

The first term in the right-hand side of equation 12 is the sum of one half of the pair-wise van der Waals (ΔE_{vw}^{ij}), electrostatic (ΔE_{el}^{ij}) and polar-solvation ($\Delta G_{GB/PB}^{ij}$) interaction energies between atoms *i* and *j* belonging to residue *res* and to any other different residue, respectively. The second term stands for the sum of the self-interaction energies of all atoms belonging to residue *res* ($\Delta G_{GB/PB}^{ii}$). Finally, the third and fourth terms represent the per-residue non-polar solvation free energy (Δ$G_{SA(res)}$) and entropy ($T\Delta S_{res}$) contributions, respectively. Δ$G_{SA(res)}$ is calculated through the ICOSA algorithm (Gohlke,Kiel *et al.* 2003); whereas $T\Delta S_{res}$ is neglected by default in the MMPBSA.py free energy decomposition protocol (Miller,McGee *et al.* 2012).

3.1.4. Approaches for residue-residue correlations calculation

The quantification of correlated motions between residues can be performed by several methods. The most popular are: (*i*) the dynamic cross-correlation (*CC*) of atomic fluctuations through the Pearson correlation calculated from covariance matrix elements (equation 13) (Hunenberger,Mark *et al.* 1995), and (*ii*) the general correlation (*GC*) coefficient coming from mutual information metric of information theory (Lange and Grubmuller 2006).

$$C_{i,j} = \frac{(r_i - \langle r_i\rangle)(r_j - \langle r_j\rangle)}{\sqrt{\left(\langle r_i^2\rangle - \langle r_i\rangle^2\right)\left(\langle r_j^2\rangle - \langle r_j\rangle^2\right)}} \quad (13)$$

In above equation, the bracket-enclosed quantities represent time-averaged values, and r_i and r_j are the positional vectors of atoms *i* and *j*, respectively. Cross correlation values encompass the range of -1 (anticorrelated) to +1 (correlated). However, this approach misses a considerable fraction of the correlated motions and, therefore, usually underestimates atomic correlations (Lange and Grubmuller 2006).

On the other hand, the mutual information ($I_{i,\, j}$) between two atoms can be determined through the equation 14, where *p(x$_i$)* and *p(x$_j$)* are the marginal distributions of x_i and x_j and *p(x$_i$,x$_j$)* is the observed joint distribution. While this method is mathematically more complex than the standard cross-correlation, it does not rely on the resulting geometries of the correlated motions.

$$I_{i,\,j} = \int\int p(x_i, x_j)\log\left(\frac{p(x_i, x_j)}{p(x_i)\,p(x_j)}\right)dx_i dx_j \tag{14}$$

Mutual information is closely-related to the concept of Shannon entropy, *H*[x], which states that the expected information content of a discrete random variable *x*, having a probability distribution *p(x)* corresponds to:

$$H[x] = -\int p(x)\ln p(x)dx \tag{15}$$

The mutual information in the case of bivariate random variables can be calculated as:

$$I[x_i, x_j] = H[x_i] + H[x_j] - H[x_i, x_j] \tag{16}$$

For an easier interpretation of the mutual information, it is thus convenient to define a generalized correlation coefficient, *GC*, as the Pearson-like coefficient derived from the mutual information. Considering collinear Gaussian distributions of unit variance, where d is the dimensionality of the data (d=3 for cartesian trajectories), the generalized correlation coefficient *GC* is defined as:

$$GC[x_i, x_j] = \{1 - e^{2I_{i,j}[x_i, x_j]/d}\}^{-1/2} \tag{17}$$

The coefficient *GC* is zero for fully uncorrelated variables and assumes values up to 1 for fully correlated variables (Lange and Grubmuller 2006).

3.1.5. Construction of residue-residue networks and calculation of preferred allosteric pathways in proteins

Protein residue networks are defined as the set of residues; i.e., the nodes, connected by edges (residue pair connections) whose length is weighted using the generalized correlation coefficient *GC* according to the equation 18 (Sethi,Eargle *et al.* 2009).

$$d_{i,j} = -\log |GC_{i,j}| \qquad (18)$$

Here, $d_{i,j}$ is the distance between contacting nodes i and j, and $GC_{i,j}$ is the pairwise correlation between them. According to this analysis, the obtained graph will produce short distances for strongly correlated residues and longer distances for residues with weak correlations.

The method of community analysis groups a large number of residues into “communities”, starting from an interaction network formed by all protein residues. However, it does not exist one single scheme for constructing the residue community network when a contact map filter is used. According to Sethi *et al.* (Sethi,Eargle *et al.* 2009), two nodes are considered connected if their heavy atoms are within a distance cutoff of 4.5 Å for at least 75% of simulation time. In recent studies, several interatomic distances were tested for obtaining the best community network filtered by contact map. In such cases, the modularity is the mainly used parameter for determining the suitability of constructed network. Nevertheless, a maximization of modularity sometimes creates unexpected community partitions. Under these circumstances, it would be much better to select the network which has the smallest number of communities and a modularity value close to the maximal one (Newman 2006).

3.1.6. Protein internal dynamics

In recent years, the calculation of distance fluctuations between Cα has become in an interesting approach to investigate the influence of each ligand on protein intrinsic dynamics, (Morra,Potestio *et al.* 2012;Vettoretti,Moroni *et al.* 2016). This methodology is based on the construction of a matrix whose elements, termed coordination propensity (*CP*) parameters, are calculated by the following equation:

$$CP_{i,j} = \langle \left(d_{i,j} - \langle d_{i,j} \rangle\right)^2 \rangle \qquad (19)$$

where $d_{i,j}$ is the distance of Cα of residues i and j in each snapshot and the brackets indicate the average distance over the trajectory.

Unlike the covariance matrix, *CP* does not depend on the selection of a particular protein reference structure, and is not affected by rotational/translational motions. For this reason, the *CP* parameter constitutes an indicative of coordination pattern in protein domains. In this sense, the residue pairs which belong to the same quasi-rigid domain will have much smaller distance fluctuations than those located in different domains (Morra,Potestio *et al.* 2012;Vettoretti,Moroni *et al.* 2016).

3.1.7. Principal component analysis

To extract the motions relevant for the function of a given protein from an MD simulation is a non-trivial task. Principal Component Analysis (PCA) or essential dynamics (ED) is a statistical method allowing the determination of uncorrelated collective coordinates, which describe the structural fluctuation of a protein during an MD simulation. Moreover, the determination of the system collective coordinates allows the identification of a small subset of dimensions or degrees of freedom accounting for the most important structural changes in a protein (the so-called essential subspace). The PCA method is based on the diagonalization of the covariance matrix built from atomic fluctuations in a trajectory from which the overall translation and rotations have been removed (Amadei,Linssen *et al.* 1993):

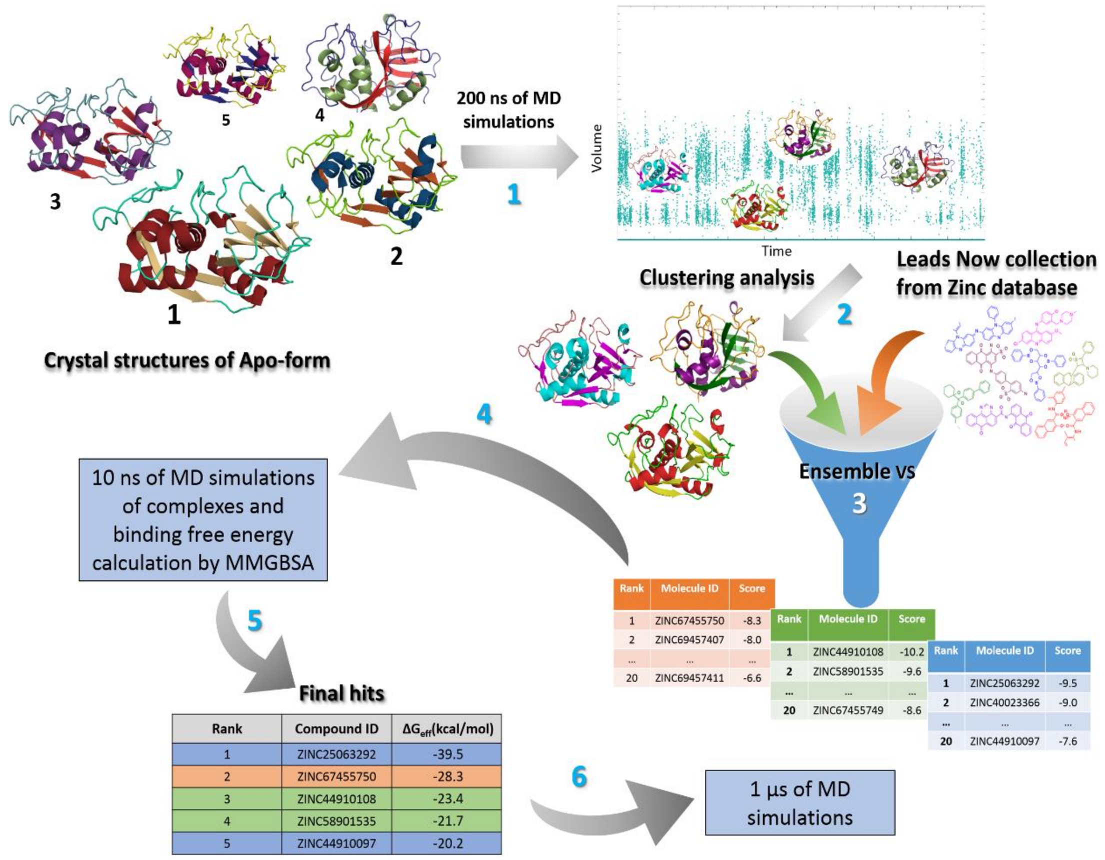
$$C_{i,j} = \langle (X_i - X_{i,0})(X_j - X_{j,0}) \rangle \tag{20}$$

in which *X* are the separate x, y, and z coordinates of the atoms fluctuating around their average positions X_0. $\langle \ldots \rangle$ represents the average values calculated from the entire trajectory. Frequently, to construct the protein covariance matrix, the Cα atom trajectory is used. Indeed, it has been shown that the Cα atom contains all the information for a reasonable description of the protein large concerted motions. Upon the covariance matrix diagonalization, a set of eigenvalues and eigenvectors is obtained. The motions along a single eigenvector correspond to concerted fluctuations of atoms. On the other hand, the eigenvalues represent the total mean square fluctuation of the system along the corresponding eigenvectors (Amadei,Linssen *et al.* 1993)

3.2. Procedures

3.2.1. Protein structures retrieval and systems setup

Twenty-five tridimensional structures of cruzain in complex with competitive inhibitors were obtained from Protein Data Bank (PDB) (Berman,Westbrook *et al.* 2000). Subsequently, all cruzain structures (include each protein chain reported in each PDB file) were compared through the calculation of pairwise Root Mean Square Deviation (RMSD) of the backbone atoms. This analysis was performed in order to find structures


3
5
4
200 ns of MD simulations
1
2
1
Crystal structures of Apo-form
Volume
Time
Clustering analysis
Leads Now collection from Zinc database
2
4
Ensemble VS
3
10 ns of MD simulations of complexes and binding free energy calculation by MMGBSA
5
Final hits
6
1 µs of MD simulations

Rank	Molecule ID	Score
1	ZINC67455750	-8.3
2	ZINC69457407	-8.0
...	...	...
20	ZINC69457411	-6.6

Rank	Molecule ID	Score
1	ZINC44910108	-10.2
2	ZINC58901535	-9.6
...	...	...
20	ZINC67455749	-8.6

Rank	Molecule ID	Score
1	ZINC25063292	-9.5
2	ZINC40023366	-9.0
...	...	...
20	ZINC44910097	-7.6

Final hits

Rank	Compound ID	ΔG_{eff}(kcal/mol)
1	ZINC25063292	-39.5
2	ZINC67455750	-28.3
3	ZINC44910108	-23.4
4	ZINC58901535	-21.7
5	ZINC44910097	-20.2

Figure 11. Flowchart of structure-based identification of cruzain allosteric inhibitors. *In the first step, five cruzain apo structures were selected to perform 200 ns of MD simulations. The second step represents the volume-based cluster analysis performed for each allosteric site, which produced three representative structures per pocket. The third step illustrates the ensemble VS accomplished for each cavity, where twenty hits were chosen per cluster (60 per allosteric site). Subsequently, 10 ns MD simulations were carried out with these 60 complexes and were employed for binding free energy calculations. In this step, a final list of five hit compounds, ranked according to the MM-GBSA results, were obtained per allosteric site. Finally, long MD simulations were conducted to analyze the mechanism of cruzain allosteric modulation by these compounds.*

with highest structural differences and to be employed in MD simulations of enzyme apo form (Figure 11, step 1). Other parameters taken into account for selection of cruzain structures were the experimental resolution in Å (<2.5 Å) and their usage in previous computational experiments, for comparison purposes.

Each ligand was removed from the selected cruzain structures. Protonation states of cruzain ionizable residues were determined at the main functional pH of the enzyme (pH=5.5) using the PDB2PQR server (Dolinsky,Nielsen *et al.* 2004). Further preparation and simulations were performed with Amber14 (Case,Babin *et al.* 2014) and AmberTools16 (D.A. Case 2017). Each system was solvated with explicit TIP3P water molecules (Jorgensen and Jenson 1998) in a cubic box whose edges were placed at a minimum distance of 10 Å from the solute surface. Subsequently, they were neutralized by replacing water molecules with Na+. Periodic boundary conditions were settled around the box. AMBER14SB (Maier,Martinez *et al.* 2015) and GAFF force-fields were selected to parametrize the protein and the organic compounds, respectively (Wang,Wolf *et al.* 2004).

Energy minimizations (EM) were performed with pmemd.MPI program of Amber14 package (Case,Babin *et al.* 2014). This procedure was conducted in two steps: (*i*) only the water molecules were minimized while restraining the solute, (*ii*) and all atoms were minimized without restraints. In both of cases, 2000 cycles of minimization were performed, the steepest descents algorithm being applied in the first 1500 cycles, and the conjugated gradient algorithm in the remaining 500 ones.

3.2.2. Molecular dynamics simulations

The equilibration procedure was conducted with pmemd.MPI and involved two sequential steps in the NVT and NPT ensembles, both keeping the solute heavy atoms restrained. During the 500 ps of NVT equilibration, the temperature was linearly increased from 10 to 300 K using the Berendsen thermostat. The subsequent 500 ps of NPT equilibration was performed at a temperature of 300 K and a pressure of 1 bar, employing the Langevin thermostat for temperature control (Schneider and Stoll 1978). The employed barostat was the isotropic position scaling method in all cases (Case,Babin *et al.* 2014). The particle mesh ewald (PME) method was employed to handle long-range electrostatic interactions (Darden,York *et al.* 1993) and the cut-off value for non-bonded interactions in all cases was set to 10 Å. Covalent bonds involving hydrogen atoms were constrained with the SHAKE algorithm (Ryckaert,Ciccotti *et al.* 1977). A time step of 2 fs was used for numerical integration of the Cartesian equation of motion and coordinate files were recorded every 20 ps.

For cruzain apo-form, 200 ns of MD simulations were performed with the five different protein crystal structures selected in pairwise-RMSD analysis. Besides, the top-scoring ligands from the Vina-score ranking list were selected for “short MD simulations” of 10 ns in complex with cruzain (Figure 11, step 4). Finally, five independent production runs of 200 ns (1µs as total) were carried out for each complex selected from the MM-GBSA results (Figure 11, step 6). The pmemd.cuda program (Salomon-Ferrer,Gotz *et al.*

2013) was employed to run the former simulations in identical conditions to those used during NPT equilibration step but without atom restraints.

3.2.3. Characterization and selection of cruzain's cavities

Initially, each allosteric site previously-predicted by the SCA method was mapped onto cruzain structure for the further analysis (Novinec,Korenc *et al.* 2014). Cruzain cavities characterization was performed employing the MD simulations performed with apo enzyme. The cavity volume should correspond to or exceed that of a ligand, the shape should enable the ligand to fit in, and the physicochemical properties should complement those of the ligand (Stank,Kokh *et al.* 2016). For this reason, we consider the cavities with a volume greater than 200 Å^3 during the simulation time of apo form to perform the VS experiments. This geometric property was calculated with POVME program (Durrant,de Oliveira *et al.* 2011) for each cavity along the simulation time. Therefore, this analysis provided an estimation of ligand accessibility for each cruzain identified pocket. In addition, the cruzain regions with more fluctuations were monitored through the calculation of per-residue Root Mean Square Fluctuation (RMSF) values. Finally, in order to characterize the electrostatic properties of each cavity, the APBS method available at PDB2PQR web-server was used to calculate the electrostatic potential on the protein surface (http://www.poissonboltzmann.org/pdb2pqr/d/web-servers) and the results were visualized with Pymol v1.6 (DeLano 2004).

To generate ten representative conformations per pocket, a volume-based clustering analysis was performed with the hierarchical-agglomerative algorithm implemented in cpptraj (Swami and Sadler 1997;Roe and Cheatham 2013). This calculation provided a non-redundant set of cruzain pocket conformations to carry out the ensemble VS (Figure 11, step 2). Furthermore, to filter these ten central structures, pairwise RMSD values of heavy atoms were calculated for each pocket, generating a symmetric matrix of 2D-RMSD. This analysis led to the identification of pocket dissimilarities, considering the residues RMSD as an additional criterion. Finally, three representative conformations were selected for each cavity (Figure 11, step 2), considering the structures with the highest differences in terms of both, volume and pairwise RMSD values, but ignoring the frequency of frame appearance in each cluster. The data of pocket volume and side chain RMSD values were normalized to keep both magnitudes comparable.

3.2.4. Virtual screening

Synthetic compounds from Lead Now library of ZINC database (http://www.zinc.org) were filtered, eliminating the potentially reactive and non-leads compounds. Subsequently, ligands were protonated at pH=5.5 and minimized employing the FF94 force field. All these previous steps were conducted with MOE platform (Chemical Computing Group ULC 2017). Finally, the compound database was converted to pdbqt format with AutoDockTools (Huey and Morris 2008).

VS against the three representative structures of selected allosteric sites (Figure 11, step 3) was carried out with AutoDock Vina v4.0 software (Trott and Olson 2010). This procedure was performed using the software default parameters, setting the number of energetically-degenerated poses to ten. During docking simulations, all rotatable bonds of each ligand were allowed to freely move around the bond axes, while the receptor structure was kept fixed. The grid box used to the screening was defined accordingly with each cavity size, using AutoDockTools (Figure 12) (Huey and Morris 2008). Twenty complexes with the highest Vina-score pose (S_{vina}) were selected per cluster, generating 60 preliminary hits for each cavity (Figure 11, step 3). Subsequently, all these complexes were used for short MD simulations in order to re-score binding poses with MM-GBSA calculation (Figure 11, step 4).

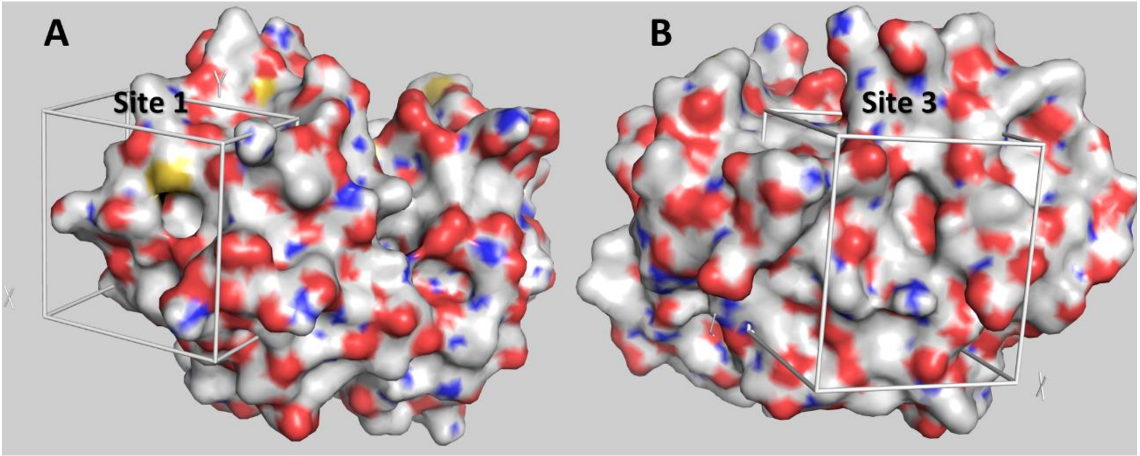


Figure 12. Structural representation of grid boxes defined with AutoDockTools for (A) site 1 and (B) site 3 of cruzain.

3.2.5. Ligand parametrization

The hit compounds derived from a ranking list of VS were optimized with Antechamber (Wang,Wang *et al.* 2006), employing the AM1-BCC method for charges calculation and the parameters of GAFF force field (Wang,Wolf *et al.* 2004). However, the ligands selected to perform long MD simulations (Figure 11, step 5) were optimized at HF/6-31G* level using Gaussian 09 package (Frisch,Trucks *et al.* 2009). In this case, electrostatic potentials (ESPs) of the geometrically-optimized structures were finally generated by single-point calculations with HF/6-31G* method and Merz-Kollman (MK) scheme (Besler,Merz *et al.* 1990). Partial atomic charges were fitted to the ESPs through the restricted electrostatic potential (RESP) method (Bayly,Cieplak *et al.* 1993). Likewise, ligand atom types, bond and dihedral angles, atomic masses and bond lengths were obtained from GAFF using Antechamber (Wang,Wang *et al.* 2006).

3.2.6. Trajectories analysis

The cpptraj module of Ambertools16 package (Roe and Cheatham 2013;Case,Babin *et al.* 2014) was used to determine trajectory parameters such as RMSD, RMSF, average structure, interatomic distances and hydrogen bonds. The calculation of RMSD values for side chain and backbone atoms was performed employing the snapshots of production run with respect to starting frame. Hydrogen bonds established between each ligand and cruzain were calculated with the default geometric definition of cpptraj. Figures of protein cavities density, tube representation of RMSF on protein structure and complexes were generated with Pymol v1.6 program (DeLano 2004).

The intramolecular interactions, i.e., hydrogen bonds, hydrophobic contacts and salt bridges were determined with Pyinteraph program (Tiberti,Invernizzi *et al.* 2014). This analysis also led to the identification of structural differences between apo and holo forms of cruzain, and the examination of crucial interactions that could affect protein stability upon ligand binding.

3.2.7. Binding free energy calculations

The ΔG_{bind} values of all protein–ligand complexes were calculated using the MMPBSA.py program of Ambertools16 (Miller,McGee *et al.* 2012). The single-trajectory approximation was used for this calculation, employing the MD simulations corresponding to the 60 complexes selected per cavity (Figure 11, step 4). The GB-neck2 model (igb=8 as a input of MMPBSA.py script) with mbondi3 radii was used for estimating ΔG_{GB} (Case,Babin *et al.* 2014). The ΔG_{bind} calculation was based on the average over all snapshots, generating a new score, which allows to obtain a re-ranked list of hit compounds (Figure 11, step 5). In the case of final selected hits, the mean ΔG_{eff}

values from simulation replicas were determined as a criterion of ligand stability into allosteric site and enthalpic strength of binding. The standard error of the mean (*SE*) was calculated by the following equation: $SE=SD/\sqrt{N}$, where *SD* is the standard deviation of the ΔG_{eff} mean values obtained from the independent simulations (*N*). Finally, per-residue effective free energy decomposition (ΔG_{res}) was carried out in order to determine the more important residues involved in cruzain-ligand interactions (Gohlke,Kiel *et al.* 2003).

3.2.8. Coordination propensity calculation

Calculation of *CP* coefficient was performed with home-made scripts following the formula reported in several works (equation 19) (Morra,Potestio *et al.* 2012;Vettoretti,Moroni *et al.* 2016). The *CP* matrices obtained from independent replicas were averaged for each system. The values obtained for the apo form were subtracted from those obtained for the holo form $\left(\langle CP_{apo}\rangle - \langle CP_{holo}\rangle\right)$ and reported for each complex.

3.2.9. Comparison of apo and holo forms through essential dynamics

Principal component analysis was performed employing Bio3d package (Skjaerven,Yao *et al.* 2014). All trajectory frames were used for this analysis in all studied systems. The overall translational and rotational motions were eliminated by fitting each trajectory to its first frame. A covariance matrix was generated using Cartesian coordinates of Cα atoms from the apo trajectory. We further projected the configurations sampled during the holo trajectories onto the two most dominant principal components obtained from PCA of apo trajectory. This analysis provides a low dimensional representation of structures facilitating the detection of subtle conformational changes.

3.2.10. Graph construction and calculations

The *GC* coefficient was calculated with g_correlation software of GROMACS package v3.2.1 (Lange and Grubmuller 2006) (Figure 13, step 2). Employing the equation 17, the *GC* of each pair of residues was obtained for each system, considering the motion of Cα atoms and averaging over independent replicas for both, apo and cruzain-bound complexes. In order to highlight the principal changes in *GC* values of analyzed systems, the matrices obtained for apo enzyme were subtracted from those obtained for holo complexes $\left(\langle GC_{holo}\rangle - \langle GC_{apo}\rangle\right)$.

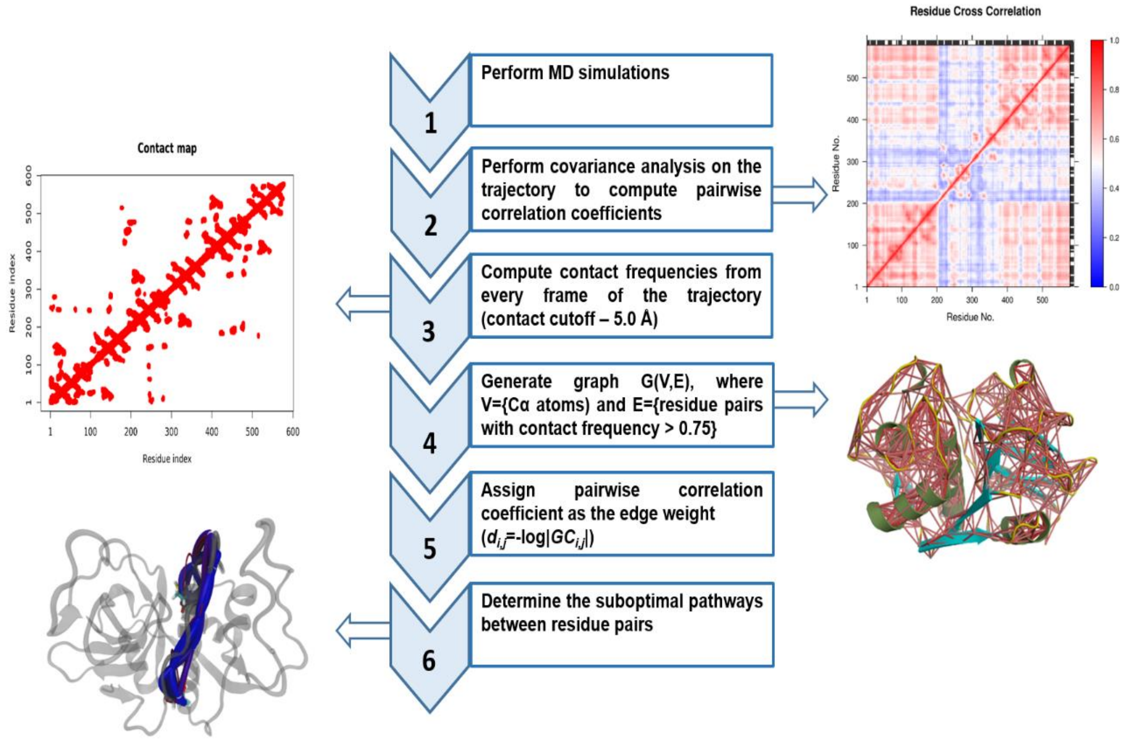

Residue Cross Correlation
Residue No.
Residue No.
Contact map
Residue index
Residue index
Perform MD simulations
1
Perform covariance analysis on the trajectory to compute pairwise correlation coefficients
2
Compute contact frequencies from every frame of the trajectory (contact cutoff – 5.0 Å)
3
Generate graph G(V,E), where V={Cα atoms) and E={residue pairs with contact frequency > 0.75}
4
Assign pairwise correlation coefficient as the edge weight ($d_{i,j}$=-log|$GC_{i,j}$|)
5
Determine the suboptimal pathways between residue pairs
6

***Figure 13. Workflow employed for construction of protein residue-residue network and calculation of allosteric suboptimal pathways**. In the first step, MD simulations of the systems are performed. In the second step, the matrix of correlation motions between residues is calculated employing the MD trajectories. The third step involves the detection of residues in contact during all trajectory with the frequency of 0.75. The fourth step is the construction of a coarse grain model of the protein, where each node represents the corresponding Cα atom and the edges are the contacts between residues. In the fifth step, the edges are weighted in function of the GC values employing the equation in bracket. Finally, the suboptimal pathways are determined between the source (residue of allosteric site) and sink (residue of orthosteric site) residues.*

In this work, the network construction and the processing of *GC* and contact map matrices were made with Bio3d package (Skjaerven,Yao *et al.* 2014) employing the methodology proposed by Sethi *et al.* (Sethi,Eargle *et al.* 2009). The distance used for contact maps calculation was 5.0 Å, as previously used in others studies (Rivalta,Sultan *et al.* 2012), in which it produced the smallest number of communities for all systems (Figure 13, step 3). The cut-off of *GC* value taken for network building was 0.5. The pairwise product between the average *GC* and contact map matrices across simulation replicates $(\langle GC_{ij}\rangle \cdot \langle Cmap_{ij}\rangle)$ was used for generating the final matrix employed in network community calculations (Figure 13, step 4). The graph community structure was built using the Girvan-Newman clustering method, which is essentially based on the edge betweenness partition criterion (Girvan and Newman 2002).The optimal paths were identified using Dijkstra's algorithm in NetworkX (Dijkstra 1959). Finally, giving a pair of residues termed source and sink, 500 suboptimal pathways were calculated employing the Weighted Implementation of Suboptimal Pathways (WISP) method (Van Wart,Durrant *et al.* 2014), and taking a residue hot spot of allosteric site and CYS25 of catalytic site as source and sink residues, respectively (Figure 13, steps 5 and 6).

4. Results and discussion

4.1. Slight dissimilarities between cruzain crystal structures

The pairwise RMSD analysis of cruzain structures shows small differences between crystal coordinates (Figure 14). The observed deviations may arise from the basal motions of the protein structures and not from large-scale movements. This fact indicates the stability of cruzain structure, notwithstanding the presence of dissimilar ligand scaffolds within its active site in the aforementioned structures. Consequently, the selection of the five structures for further structural analysis was based mainly on the following conditions: (*i*) relatively-large pairwise RMSD values, (*ii*) different experimental sources and (*iii*) high resolution (<2.5 Å).

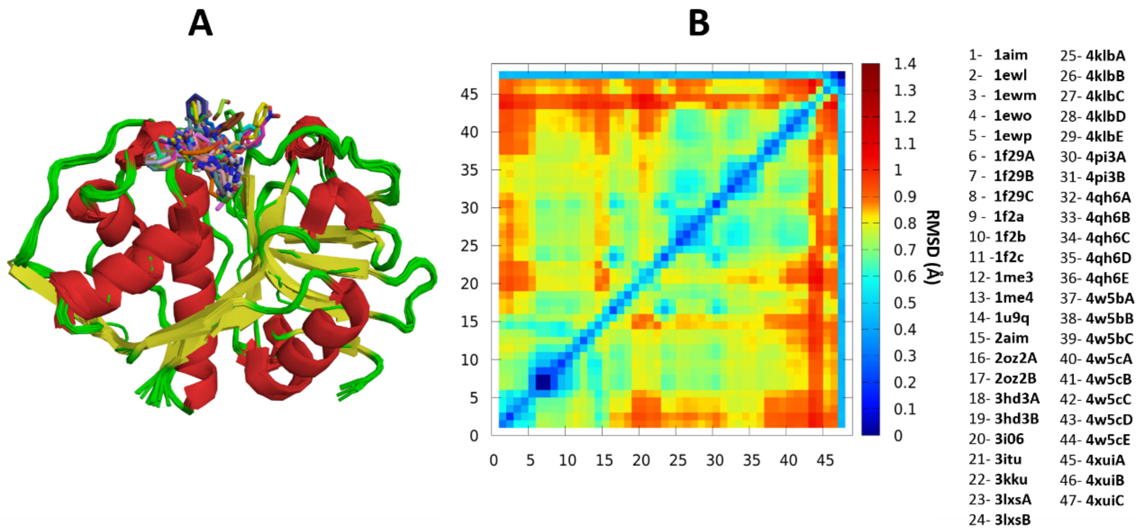


Figure 14. Comparison of crystal structures reported for cruzain*. (A) Superimposed structures of cruzain obtained from PDB database. Secondary structure elements are colored as follows: alpha helices (red), beta sheets (yellow) and loops/turn (green). All ligands are positioned in the active site. (B) Pairwise RMSD calculated for the backbone atoms of cruzain crystal structures (47 in total considering each cruzain copy solved within the same PDB file). PDBID of each analyzed structure is specified in alphabetical order on the right hand side.*

In the 2D-RMSD matrix (Figure 14B), the structures 1-5, 16-22 and 35-47 have the greatest pairwise RMSD values. In the 1-5 range, all the structures, except for 1aim, were solved in the same study. From the previous structures, 1ewp was selected because it has the best resolution (1.75 Å). In the 16-22 range, 3iut and 3i06 were selected since they have the best resolution values (1.2 Å and 1.1 Å respectively) and were determined in different studies. In addition, 1aim and 2aim were determined in the same work, but we selected 2aim because it has the greatest pairwise RMSD. Finally, 1me4 structure was chosen given its usage in previous simulations conducted by Durrant *et al.* (Durrant,Keranen *et al.* 2010). Structures corresponding to 37-47 range were discarded because their low resolution.

4.2. Dynamical details of cruzain allosteric sites reveal transient pockets and functional gate in a conserved groove of papain-like cysteine proteases

After mapping the putative allosteric sites of HCatK onto the surface of cruzain through structural alignment (Figure 15), the characterization of cruzain cavities was performed through a systematic comparative analysis of enzyme conformational dynamics. The results show that site 1 and site 3 display a suitable volume size along simulation time (Figure 16, data not shown for subsites with volume values lower than 200 Å^3); therefore, they were chosen for further VS experiments. Subsequently, three central structures of different clusters, which satisfy the condition exposed in section 3 (see Methodology), were selected per site (Figure 17). The results of pairwise RMSD and

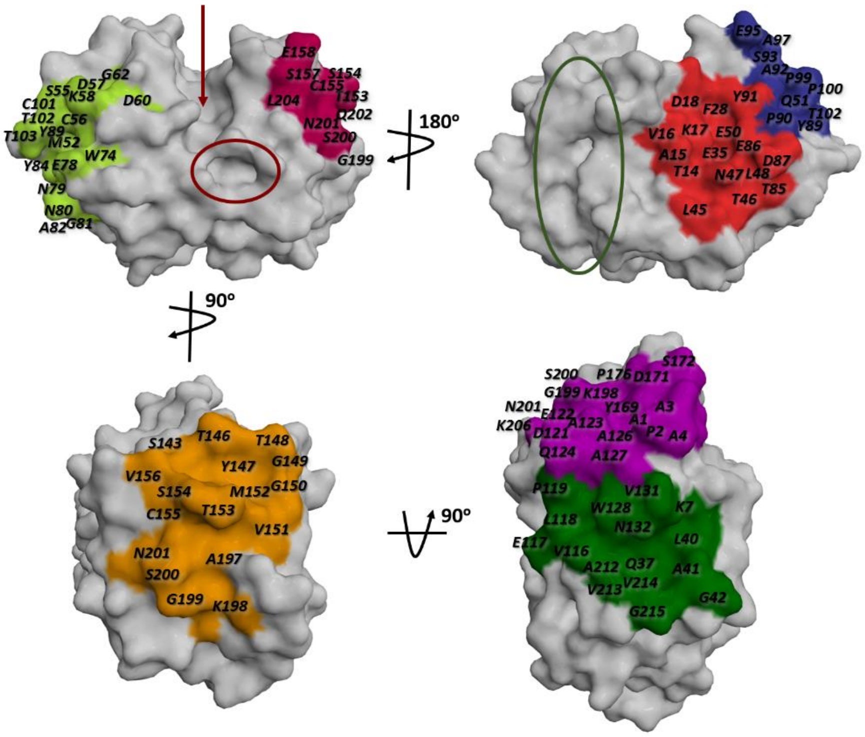


Figure 15. Residue composition of previously-predicted allosteric sites of papain like-cysteine proteases. *Cruzain potential allosteric pockets corresponding to predictions performed by of Novinec et al. (Novinec,Korenc et al. 2014) for papain-like cysteine proteases superfamily. These sites are defined from one to seven, i. e., site 1 (lemon), site 2 (pink), site 3 (red), site 4 (blue), site 5 (orange), site 6 (violet) and site 7 (green). The position of the active site is pointed with an arrow and adjacent subsite identified by Durrant et al. (Durrant,Keranen et al. 2010) is marked with red circle. The site 0 (glycosaminoglycan binding site in HCatK) is highlighted with a green circle.*

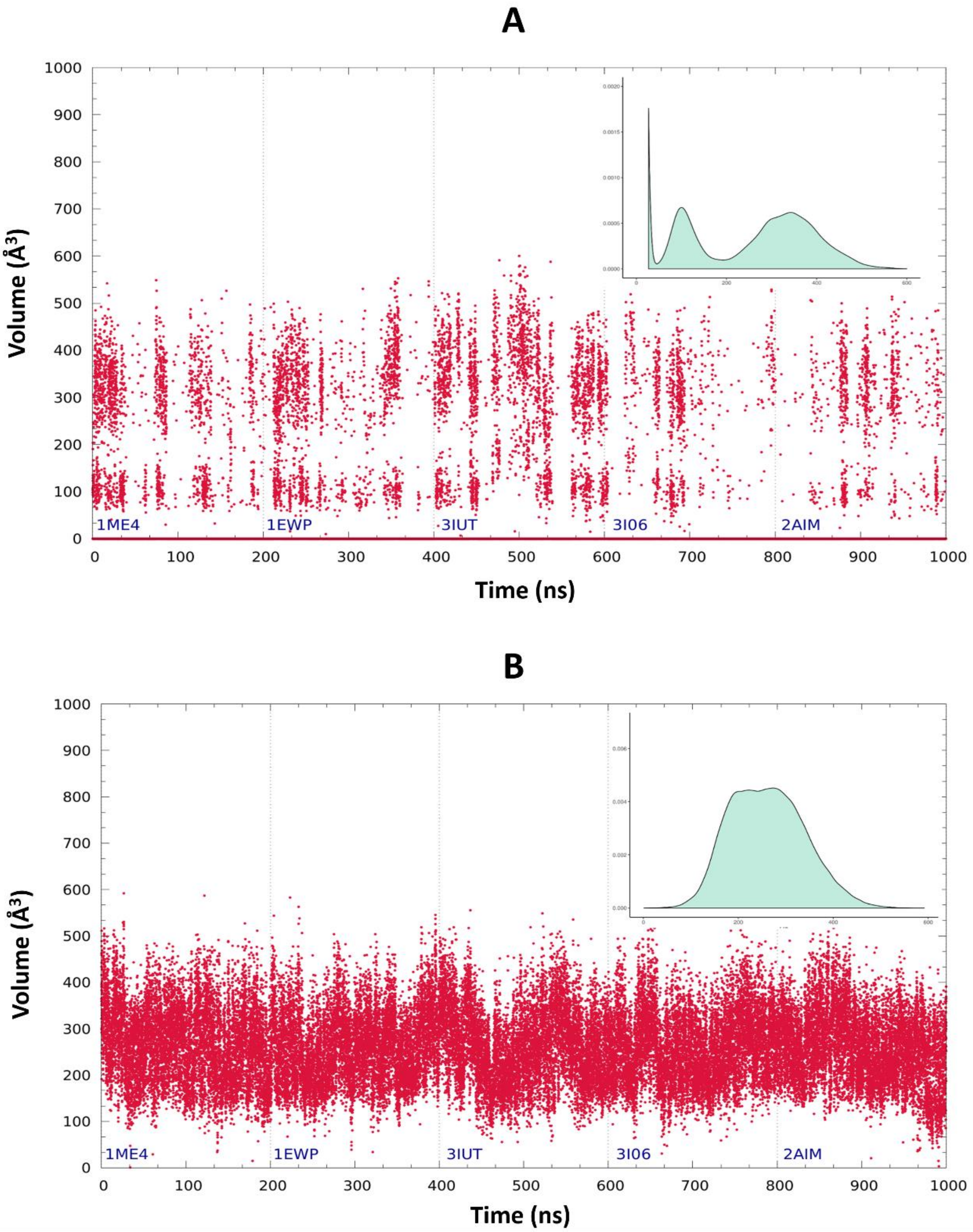


__Figure 16. Dynamical analysis of cruzain site 1 and site 3.__ Time evolution of instantaneous volume values for site 1 (A) and site 3 (B). Graphs were subdivided according to the cruzain structure employed in each simulation (PDBID). Insets within each graph correspond to the probability histograms of volume along the MD trajectory.

volume comparison between all generated central structures of site 1 and site 3 are shown in Figure 18. The time evolution of cavity volume stability indicates that site 3 is more stable than site 1 (Figure 16). However, the later site reaches higher volume values because it is formed by a flexible fragment of cruzain loop, which is homologue of cathepsin B occluding loop (loop$_{88-109}$) (Figure 19A). On the other hand, the distribution

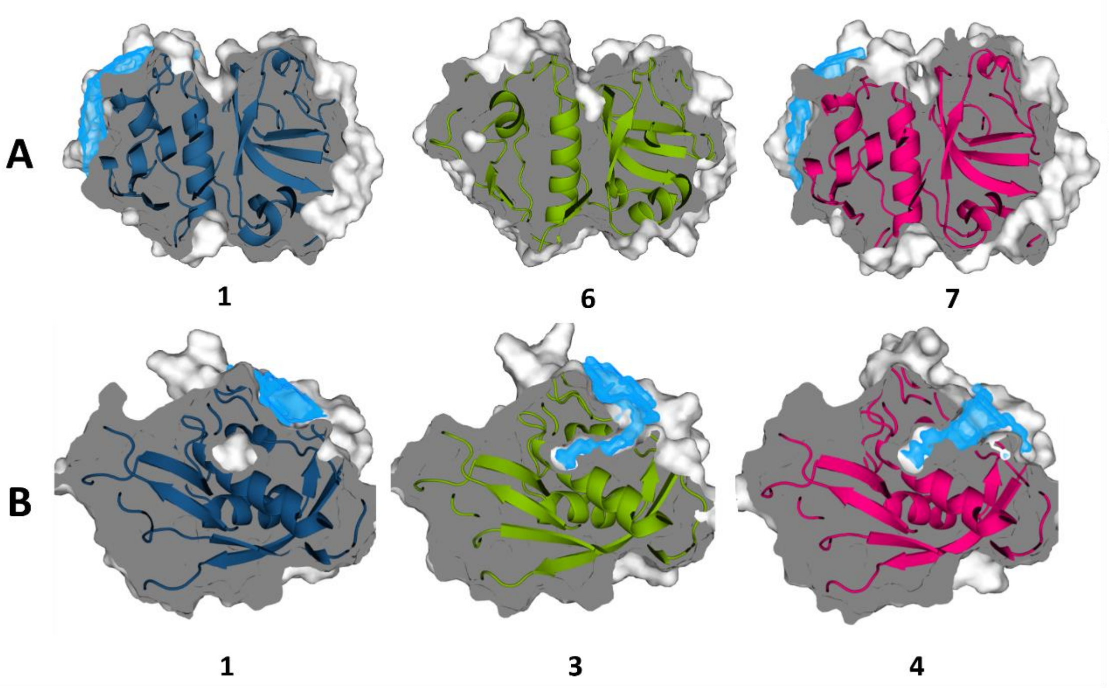


Figure 17. Structural representation of selected clusters of cruzain site1 and site 3 employed in VS experiments. *Representative structures of (A) site 1 and (B) site 3. Volume density of three selected representative structures is colored in blue and the cluster identifier is depicted in each case.*

of conformational space of site 1 shows three well-defined populations, being the structures with zero volume the most representative ones (top right of Figure 16A and Figure 17A). These predictions suggest that site 1 constitutes a cruzain transient pocket. Remarkably, transient pockets have been suggested as potential allosteric sites in proteins (De Vivo,Masetti *et al.* 2016;Stank,Kokh *et al.* 2016). Additionally, the main residues of this site are non-conserved (Figure 19A) (Durrant,Keranen *et al.* 2010), which makes it suitable for the design allosteric inhibitors.

On the other hand, cruzain site 3 showed a conformational space characterized by two overlapping normal distributions of volume values (Figure 16B). However, there is an intermediate representative structure (cluster 3) that shares structural characteristics with the remaining two (Figure 17B). The occurrence of this leads to extensive overlap of volume distribution histograms. These results illustrate an occlusion of site 3 internal cavity (cluster 1, Figure 17B), in which the salt bridge established between conserved residues LYS17 and GLU35 (Figure 19B) acts as “functional gate” of this groove. It is important to highlight that the “closed form” of site 3 is the most prevalent in crystal structures reported for cruzain, since generally, LYS17 and GLU35 are at optimal distance to form this salt bridge (~3.5 Å in crystal structures). Interestedly, two non-

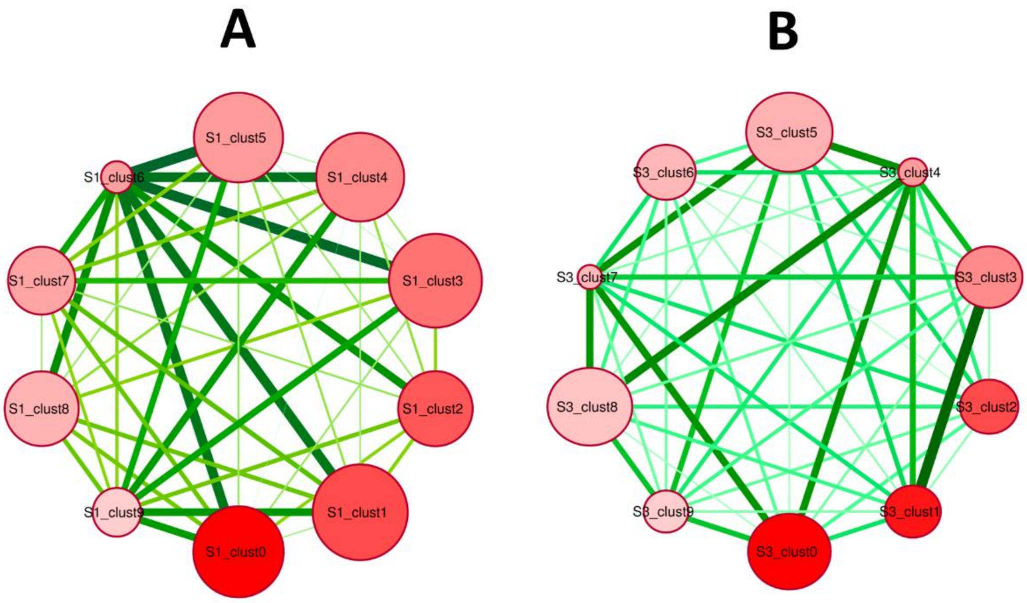


Figure 18. Comparison of clusters generated for each site by volume-based cluster analysis. *(A) Clusters (clust) calculated for site 1 (S1) and (B) site 3 (S3). Node size is proportional to volume value and color gradient correspond to the frames number conforming each cluster. Edge color and thickness are proportional to pairwise RMSD values.*

conserved glutamic residues, predicted in the protonated state, are present in the vicinity of LYS17 and GLU35 (Figure 19B). Interestingly, the protonation states of this patch of ionizable residues at the protein surface is likely to be highly susceptible to pH variations, which, in turn, may have important functional implications. This phenomenon cannot be observed through conventional MD simulations techniques, but in theory, it may affect the "open-close" dynamical equilibrium of this groove.

The hypothesis that site 3 is involved in allosteric regulation and/or enzyme structural stability was previously proposed by Durrant *et al* (Durrant,Keranen *et al.* 2010). In that work, the authors emphasized that the presence of two conserved residues lying outside site 3, i.e., SER49 and GLN51, may stabilize this groove within papain-like proteases. In addition, they underlined the role of conserved residues TYR89 and PRO90 in the rigidity of loop$_{88\text{-}109}$ and the establishment of hydrogen bond network involving residues 47, 86 and 91. All these findings point out the function of site 3 in internal structure stabilization of cruzain (Durrant,Keranen *et al.* 2010).

The RMSF analysis shows the same fluctuation pattern in the five replicas of the apo-form (Figure 20B, C). Here, we observed the main flexibility in cruzain loops, highlighting the region corresponding to loop$_{88\text{-}109}$. These evidences explain the huge

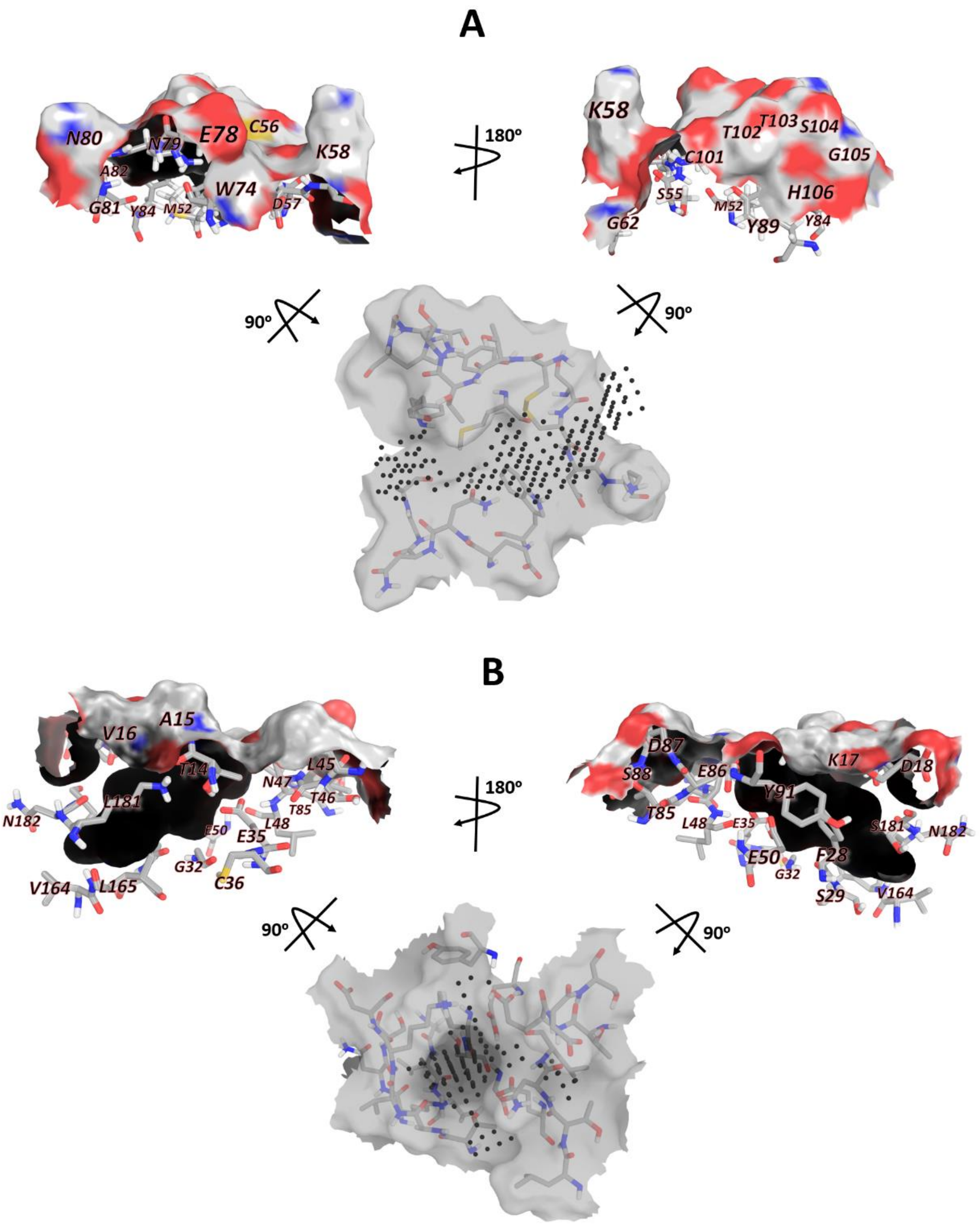


Figure 19. Residue composition of cruzain site 1 and site 3. *Surface representation of the two opposite orientations and top view of cruzain (A) site 1 and (B) site 3. Protein surface is colored according to atom type and volume size is depicted as a grid of equidistant points.*

volume variation of site 1, since several residues of this region are assembling this subsite (Figure 19A). Moreover, RMSD values calculated for heavy atoms showed different time evolution behavior between site 1 and site 3 (Figure 20A). In this sense, site 3 showed relatively stable RMSD values, whereas site 1 displayed structural fluctuations during simulation time, which was expected for a site that displays volume instability.

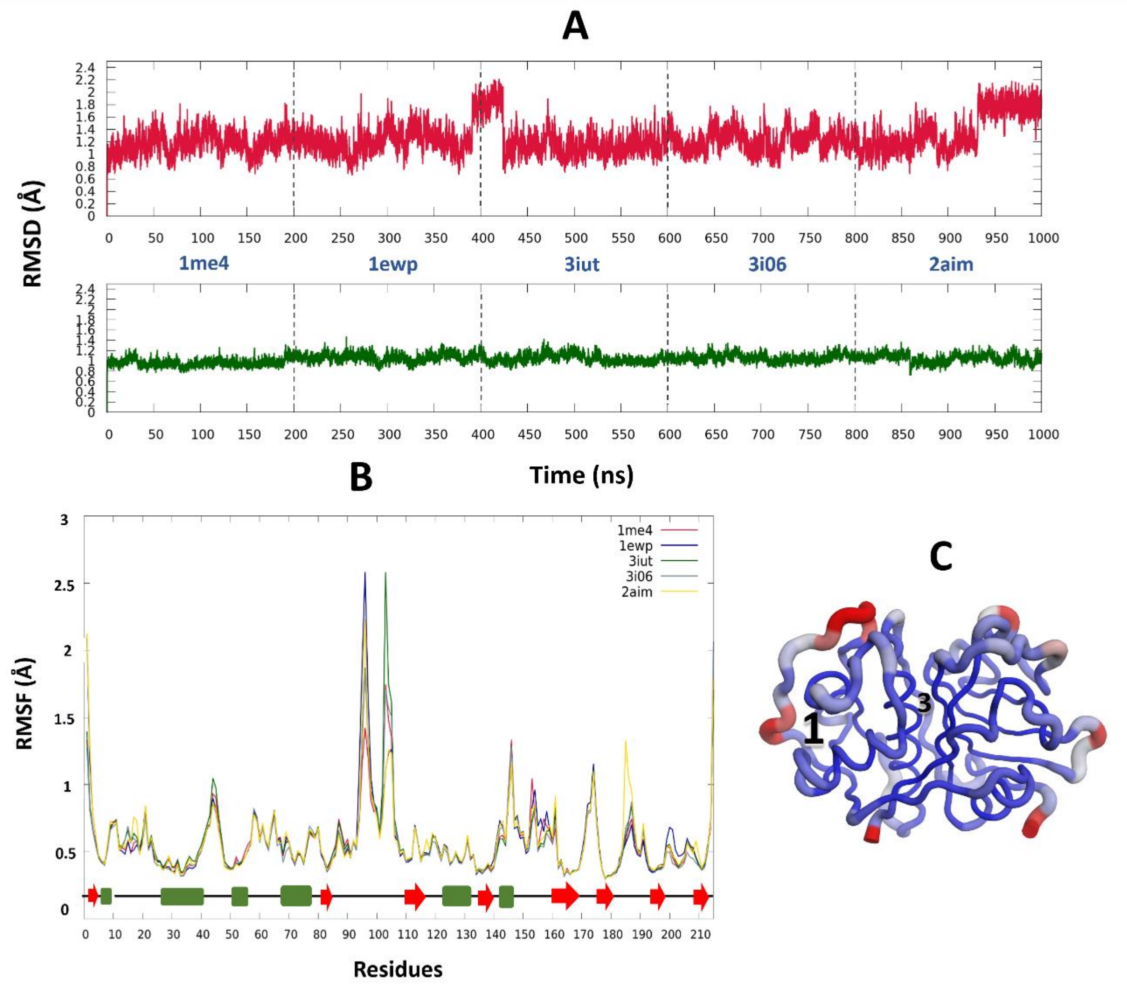
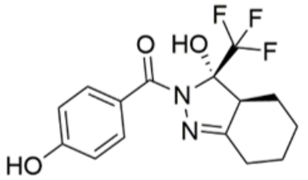
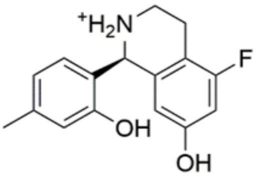
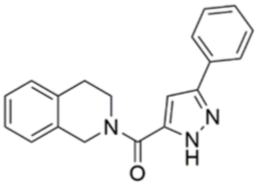
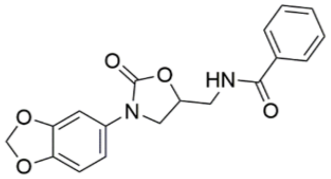
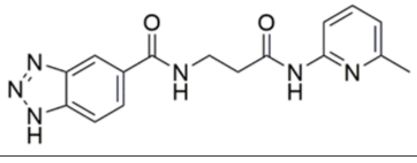


Figure 20. Allosteric sites stability in apo form of cruzain. *(A) Time evolution of RMSD values calculated for heavy atoms of site 1 (red) and site 3 (green). Each replica was labeled with the PDBID of the cruzain structure employed as starting structure in each case. (B) RMSF calculated for each replica of cruzain apo form. (C) Tube representation of free enzyme average conformation. Tube size is proportional to the per-residue atomic fluctuations computed for C-alpha atoms and the regions with high fluctuations are colored in red.*

4.3. Preferential binding of identified hits to allosteric site 3 of cruzain

According to the methodology presented in section 3, the prediction of allosteric modulators of cruzain was based on two rounds of high-throughput scoring: (*i*) Autodock Vina scoring function, and (*ii*) post-processing with MM-GBSA calculations. The list of compound ID generated in the ensemble VS, as well as their respect energy scores obtained from Autodock Vina and MM-GBSA calculation, are listed in Tables A1 and A2 for site 3 and site 1, respectively. Interestingly, these results show non-overlapping sets of selected compounds corresponding to different structural clusters of the same cavity. This underscores the importance of using different receptor conformations in VS experiments to increase the number of compound scaffolds. In addition, compounds having the lowest energy values among all sets were detected in VS against the cluster central structures that possess the largest volumes (Figure 17, Tables A1 and A2). This reinforces the necessity of exploring the protein conformational space in order to identify different pocket structures. In our case, despite having a large number of cruzain structures (25), it was necessary to perform MD simulations to observe these conformational changes prior to screening execution.

Table 1. Compound ID, MM-GBSA results, S_{vina} and chemical structure of top five hits. [a]

Rank	Compound ID	ΔG_{eff} [b]	S_{vina}	Chemical structure (pH=5.5)
1	ZINC00352089[c]	-41.04	-9.10	
2	ZINC83627668[c]	-31.02	-9.10	
3	ZINC17322062[c]	-25.21	-9.40	
4	ZINC09120852[d]	-24.52	-6.60	
5	ZINC58209662[c]	-23.81	-7.80	

[a] All energies are in kcal/mol

[b] Effective free energy $\Delta G_{eff} = \Delta E_{MM} + \Delta G_{sol}$

[c] Compounds derived from VS against site 3

[d] Compounds derived from VS against site 1

Some compounds with lowest values of MM-GBSA energy were selected as final

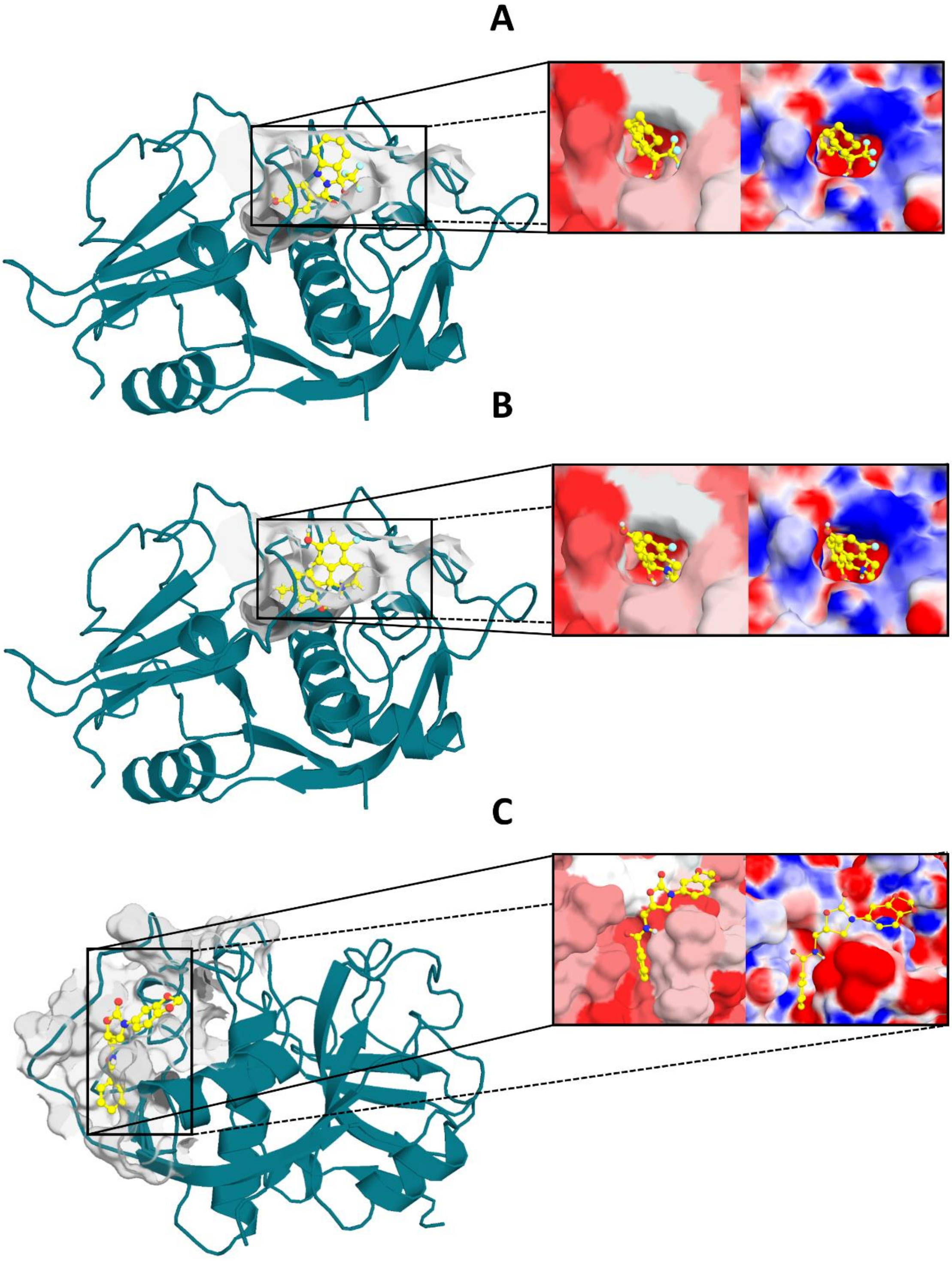


Figure 21. Selected docking poses for the best hits of each pocket. *Top scoring poses of (A) ZIN00352089 and (B) ZINC83627668 docked within site 3 pocket, and (C) top scoring pose ZINC09120852 docked within site 1. Protein surface representation is colored according to hydrophobicity and electrostatic potential. The color scale for hydrophobicity goes from red (hydrophobic) to white (hydrophilic), and for electrostatic potential, it ranges from red (electronegative potential) to blue (electropositive potential).*

hits to perform long MD simulations and to study a possible allosteric mechanism exerted in both sites. In Table 1, the top five compounds are listed and ranked according to the values obtained in MM-GBSA calculations. Note that only the compound ZINC09120852 is derived from VS against site 1. This is a consequence of large solvent accessible area present in site 1 (Figure 19A), where the interactions of ligands with solvent molecules are likely to destabilize ligand-cavity interactions in short MD simulations (10 ns). Conversely, site 3 is a buried groove, which can establish more stable interactions with ligands, but, again, at expense of large desolvation of the ligand. The ΔG_{eff} values of the compound list lie within -41 to -23 kcal/mol range (Table 1). The discarded compounds having ΔG_{eff} values around -23 kcal/mol could be included in subsequent *in silico* and/or experimental studies (see Tables A1 and A2).

Finally, the first two compounds of site 3 and the only one of site 1 were selected for long simulations in order to analyze the allosteric effects triggered by their binding to cruzain. In Figure 21, the top scoring docking poses of these ligands are represented. Moreover, each pocket surface was colored according to the electrostatic potential and hydrophobicity. These results show that site 1 is composed by residues with diverse electrostatic and hydrophobicity properties (Figure 21C), allowing to obtain compounds scaffolds that possess polar and non-polar interacting groups such as ZINC09120852 (Figure 21C). On the other hand, site 3 has a positive region at the entrance, in contrast with its internal area, which is highly negative. This is due to the presence of LYS17 in more solvent-exposed part of site 3 and glutamic acids (GLU35, GLU50 and GLU86) positioned in the pocket inner region. The two top-scoring ligands selected for this site (ZINC00352089 and ZINC83627668) contain aromatic rings placed within the pocket hydrophobic region (Figure 21A, B). In addition, the polar substituents of these ligands may interact with polar regions through hydrogen bonds and/or salt bridges in the case of ZINC83627668 (positive charged, see Table 1).

A concrete evidence of allosteric modulation in papain-like protease family, mediated by ligand binding to site 1 and 3, is a previous study reporting several allosteric regulators of HCatK (Novinec,Lenarcic *et al.* 2014). In that work, various allosteric inhibitors were predicted as compounds binding the sites 1 and 3 of the human protease. However, none of these previously identified compounds were predicted to bind the narrow groove of site 3 (Novinec,Lenarcic *et al.* 2014).

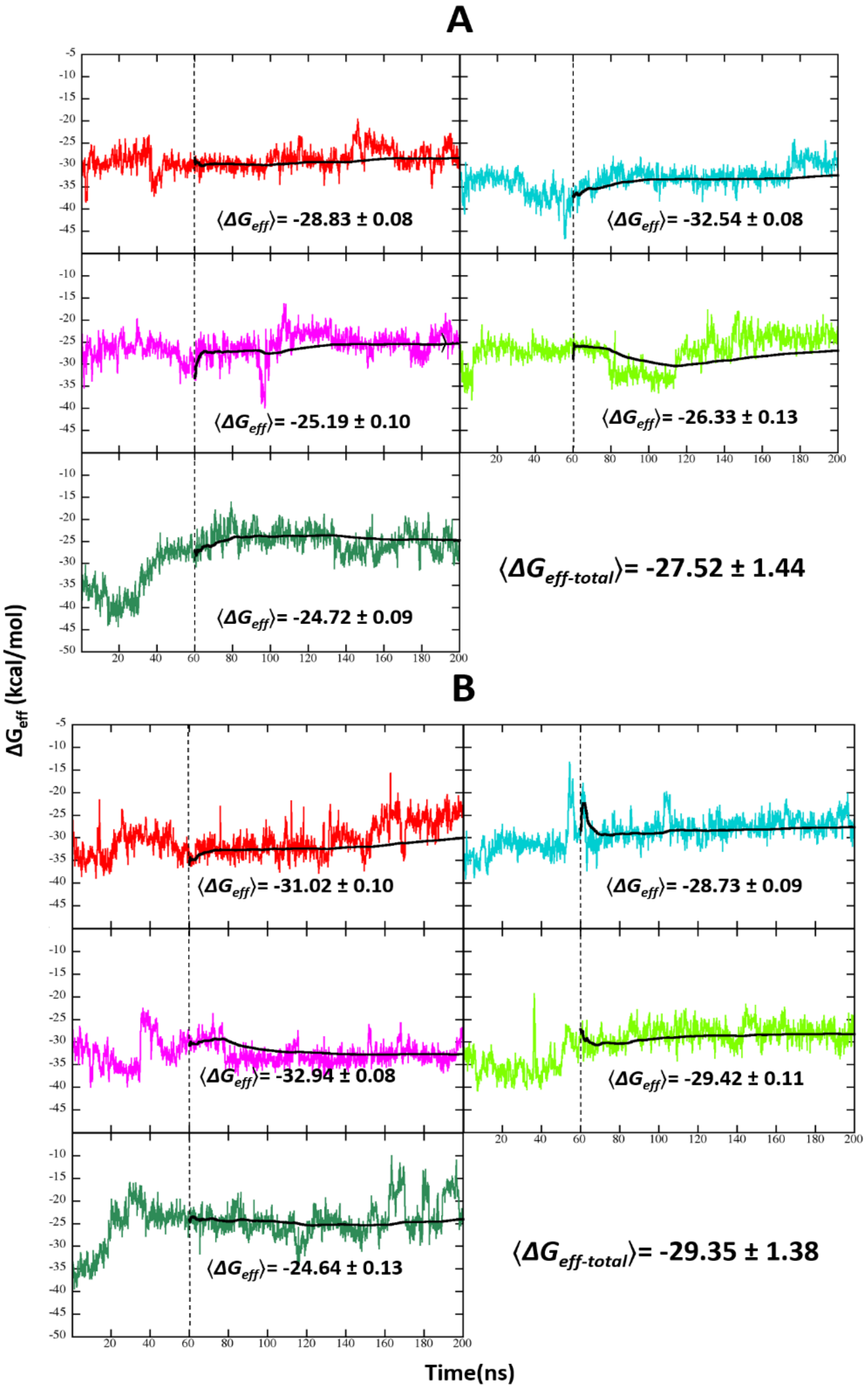


Figure 22. Time evolution of instantaneous ΔG_{eff} values for cruzain complexes. *Effective binding free energies of (A) cruzain-ZINC00352089 and (B) cruzain-ZINC83627668 complexes are shown together with accumulated mean values (black). Dashed lines indicate the MD equilibration time of complexes simulations. Every complex was labeled with the corresponding ΔG_{eff} value and average value with standard error of mean over five replicas was reported in right bottom of each graph.*

4.4. Selected hits display persistent interactions with site 3 during 1 µs of simulation time

The stability and convergence of long MD simulations performed for cruzain complexes were monitored through per-frame ΔG_{eff} and ligand RMSD (Figure 22 and

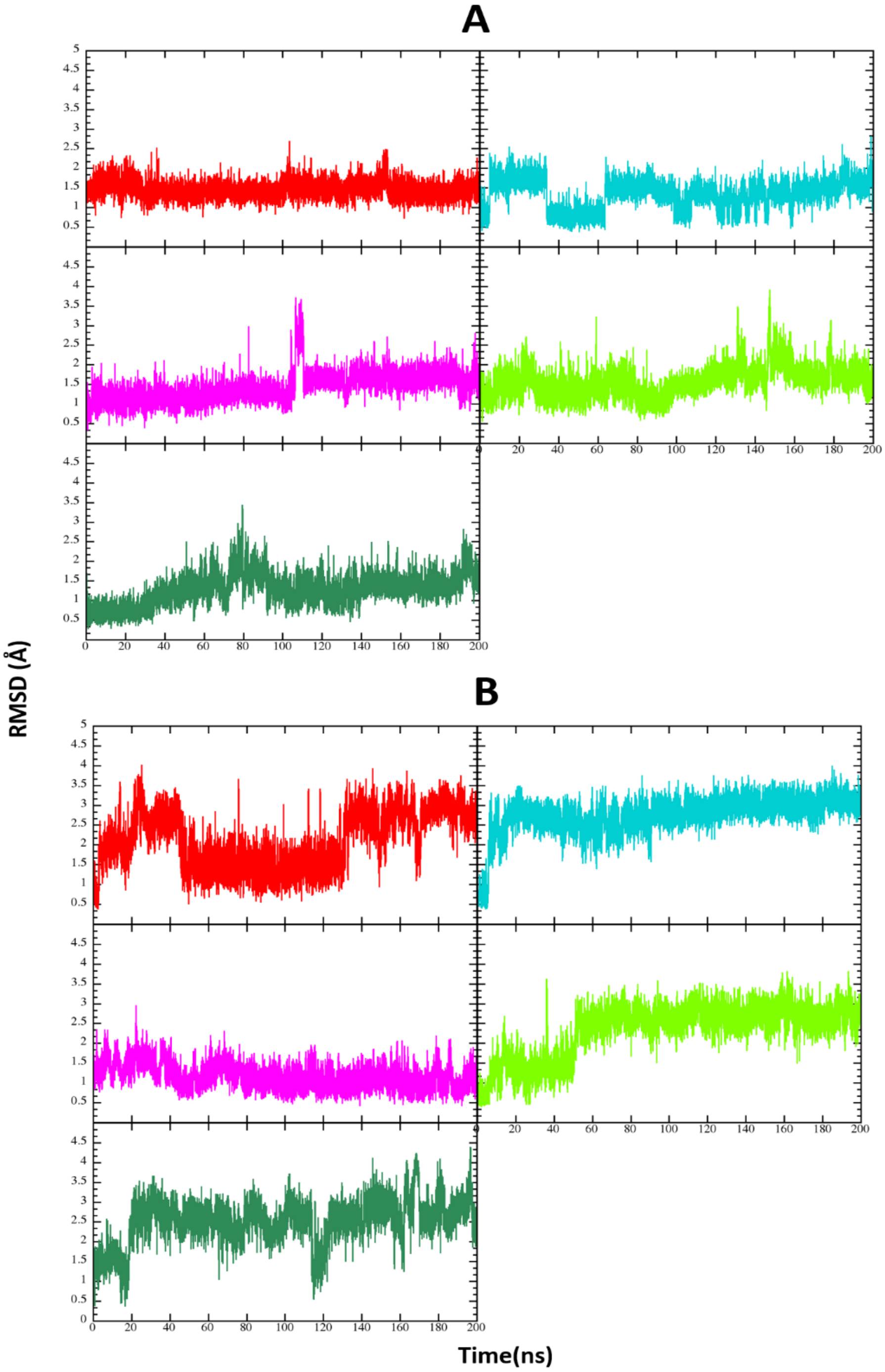


Figure 23. Ligand instability along simulation time. *RMSD time profiles calculated with respect to the initial frame of MD simulations corresponding to (A) ZINC00352089 and (B) ZINC83627668 ligands. Different colors represent an individual replica in each case.*

Figure 23). ΔG_{eff} profiles for complexes formed in site 3 are shown together with accumulated mean values. Note that ΔG_{eff} values are variable for each replica in both analyzed systems. After checking the structures of each replica, we concluded, however, that these variations are not caused by appreciable structural changes at the protein-ligand interface. Therefore, they are likely to arise from transient stability changes in the interactions formed at the interface in each replica. Additionally, the accumulated mean values become stable in most cases, except for two replicas of ZINC00352089 complex (Figure 22A). The equilibration time of complexes was selected according to the behavior of instantaneous ΔG_{eff} values across the simulation replicas. Accordingly, 60 ns was chosen as the start point for the rest of MD simulations analysis (Figure 22). On the other hand, two replicas were simulated for ligand bound to site 1, however, a complex dissociation was observed after 50 ns of simulation time (data not shown). The latter points out the lack of a real affinity of the ligand for this site. Finally, the absolute ΔG_{eff} values were calculated as the average of five replicas in order to stablish the best hit of studied compounds. The results show subtle differences between both compounds (±2 kcal/mol), therefore, both ligands may have the same priority in terms of enthalpic contribution for further experimental assays.

4.5. Aliphatic chains of site 3 display large energy contribution to ligand binding

In order to get insights into ligand interactions with cruzain site 3, the MM-GBSA approach was employed to decompose the effective binding free energy into per-residue contributions (Figure 24). In general, we observed that despite site 3 surface being highly charged, van der Waals interactions have the main role in cruzain-ligand interfaces. This is because the polar desolvation energy is most unfavorable in residues which stablish polar interactions, i. e., GLU35, GLU86 and LYS181. In both compounds the major hot spot residue is LYS17, which interacts through its aliphatic side chain, exposing its charged amine group to the solvent. Also, THR14, PHE28, ASN47 and GLU50 are present as critical interacting residues in both analyzed systems.

The decomposition of ΔG_{res} in side chain and backbone contributions led to the identification of critical positions within site 3. In almost all residues represented in Figure 24, the energy contributions of side chain are larger than those of the backbone, except for ASN182 and LEU48. This evidence suggests the essential role of residue positions conforming site 3 in the interaction with ligands. From a point of view of allostery implications, three energetically-relevant residues, i. e., THR14, LYS17 and PHE28, are conserved positions within papain-like cysteine proteases (Durrant,Keranen *et al.* 2010), which also were characterized through SCA method as residues belonging to functional sectors, and are likely to participate in allosteric regulation of HCatK (Novinec,Korenc *et al.* 2014). Otherwise, GLU50 is a semi-conserved position, also lying within the same sector of former residues, which forms hydrogen bonds with both ligands

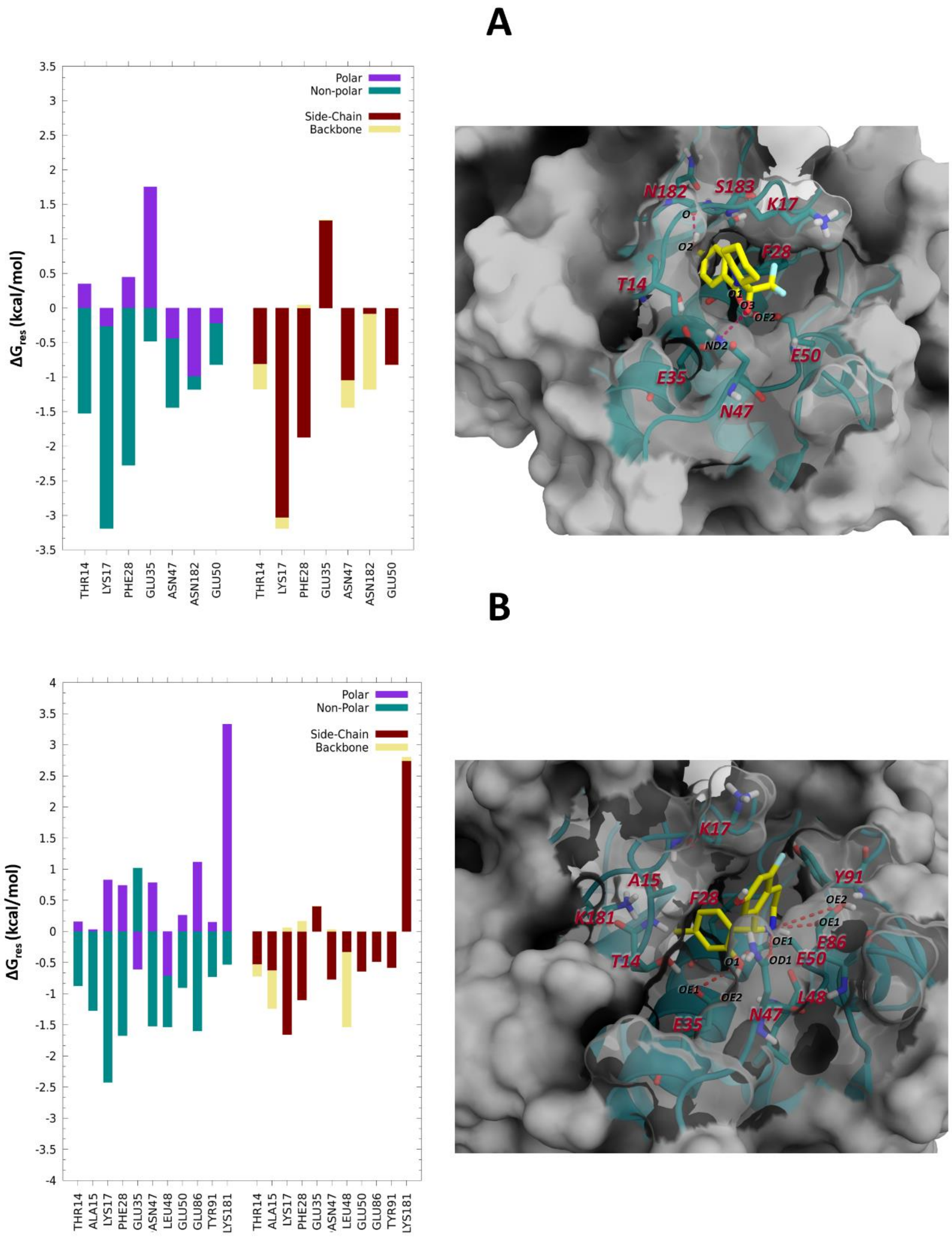


Figure 24. Per-residue free energy decomposition of cruzain complexes. *The side chain, backbone, polar and non-polar contributions of each residue is displayed for (A) cruzain-ZINC00352089 and (B) cruzain-ZINC83627668 complexes. A structural representation of each interface is depicted where residues considered hot spots are labeled together with the name of atoms forming the main inter-molecular hydrogen bonds (red dashed lines).*

(Table 2). In addition, ASN47 is a non-conserved residue that establishes hydrogen bonds with both ligands, which may be crucial for the selectivity of designed cruzain allosteric inhibitors (Figure 24 and Table 2). Note the unfavorable total contribution of conserved

GLU35 in both complexes stability, where the polar desolvation energy is more unfavorable than its contribution forming hydrogen bonds (Figure 24).

In the cruzain-ZINC00352089 complex, ASN182 establishes the more stable hydrogen bond through its backbone oxygen (Figure 24A and Table 2). Interestingly, ASN182 is a highly conserved residue whose participation in catalytic mechanism of cruzain has been proposed recently (Arafet,Ferrer *et al.* 2017). In the cruzain-ZINC83627668 complex, GLU86 also forms a hydrogen bond through its carboxylic oxygen atoms, but its polar desolvation energy is more unfavorable because the presence of positive charge on compound N atom. In addition, ZINC83627668 establishes important van der Waals contacts with THR14, ALA15, PHE28, GLU86, GLU50, LEU48 and TYR91. However, LYS181 displays a more negative contribution to cruzain-ZINC83627668 binding process because of the high cost of amine group desolvation, which is facing the phenolic ring of ZINC83627668 (Figure 24B).

Table 2. Principal hydrogen bonds established along simulation time between cruzain allosteric site and hit compounds.

System	Acceptor	Donor	Stability (%)	AvgDist (Å)
Cruzain-ZINC00352089	ZINC00352089_O1	GLU50_OE2	68.41	2.71
	ASN182_O	ZINC00352089_O2	44.96	2.75
	ZINC00352089_O3	ASN47_ND2	21.03	2.88
	ZINC00352089_O3	ASN47_ND2	16.25	2.88
Cruzain-ZINC83627668	GLU35_OE2	ZINC83627668_O1	56.01	2.56
	GLU35_OE1	ZINC83627668_O1	40.86	2.57
	GLU50_OE1	ZINC83627668_N1	37.18	2.77
	GLU86_OE1	ZINC83627668_N1	32.02	2.85
	ZINC83627668_O1	GLU50_OE2	25.75	2.78
	ASN47_OD1	ZINC83627668_N1	16.05	2.83

4.6. Compound ZINC83627668 increases the flexibility of cruzain structure

Subsequently, we turned the attention to the characterization of effects on cruzain internal dynamics after ligand binding to site 3. The Coordination Propensity (*CP*) defined in section 3 was calculated for each system in order to characterize their global rigidity/flexibility patterns. Here, we intend to seek changes produced by ligand binding in the coordination of amino acids belonging to different protein domains, which could

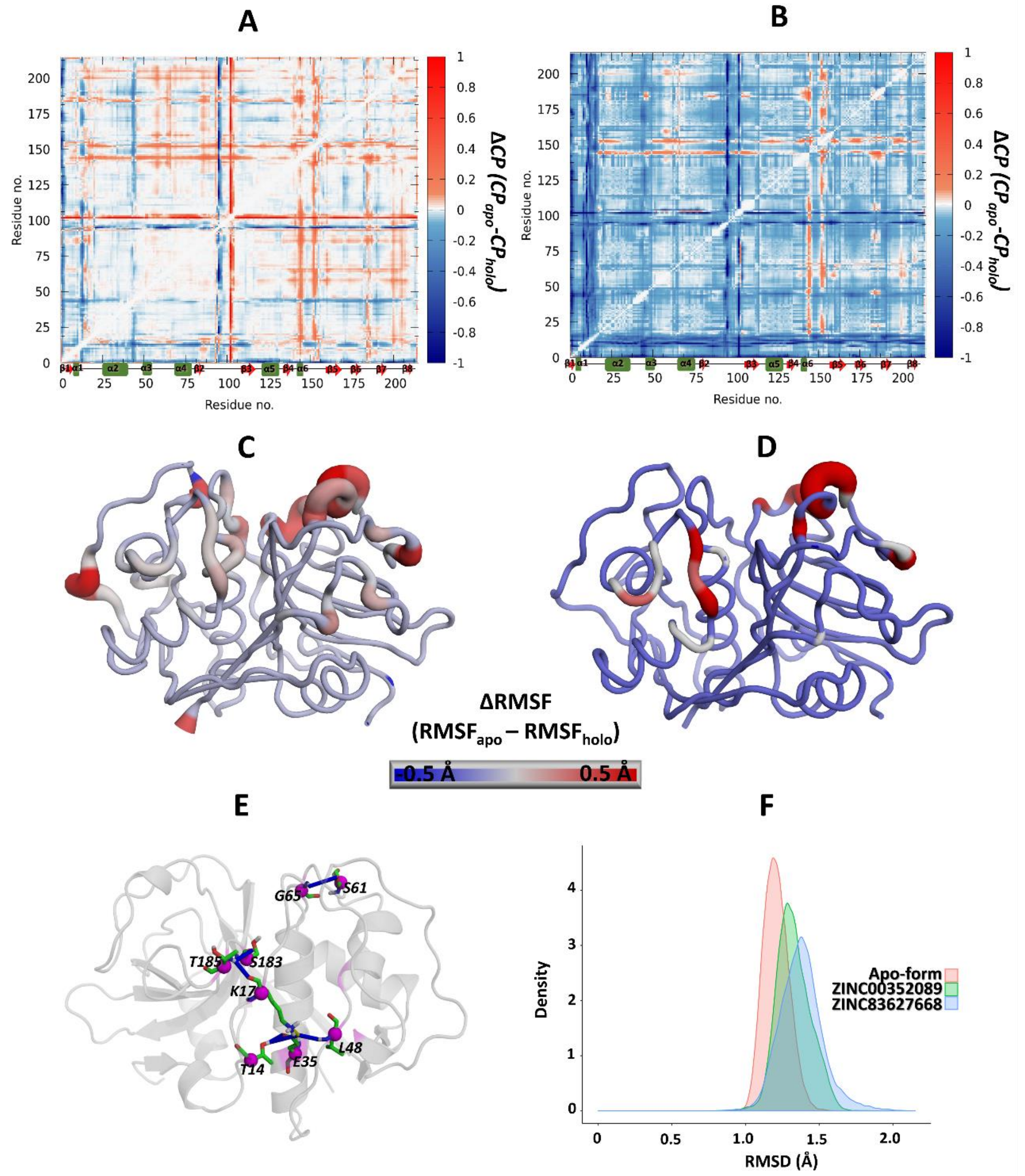


***Figure 25. Analysis of motions of the apo and holo forms of cruzain**. Graphic representation of matrices corresponding to CP differences of (A) cruzain-ZINC00352089 and (B) cruzain-ZINC83627668 complexes with respect to the apo form. The matrix scale is colored from blue (higher flexibility in holo form) to red (higher flexibility in apo form). In addition, differential RMSF is represented for (C) cruzain-ZINC00352089 and (D) cruzain-ZINC83627668 complexes. All differences were calculated by subtracting the values of holo state from those of apo state. (E) Intra-molecular interactions which were broken in both holo forms. (F) RMSD distributions corresponding to three analyzed systems which were calculated for backbone atoms respect to the same initial structure.*

be reflected as changes in rigid dynamics of inter-domains. The matrices obtained from the *CP* differences between apo and holo forms are shown in Figure 25A and B. Positive values of in the matrices (red color) indicate higher flexibility in the apo form, while negative ones (blue color) denote higher flexibility in the holo form.

The large white region of matrix corresponding to cruzain-ZINC00352089 complex indicates similar patterns in flexibility/rigidity between apo and holo forms. The main differences are allocated in loops regions, while the same degree of internal coordination is maintained in residues belonging to secondary structures (Figure 25A). However, allostery has been observed in the absence of large-scale conformational changes, suggesting that subtle changes in protein dynamics can induce a population shift in conformational ensemble without substantially changing the mean conformation of the protein (Tsai,del Sol *et al.* 2008;Nussinov and Tsai 2015). In this system, the greatest coordination differences are noticeable in loop$_{84-109}$ and loop$_{139-161}$, where differential patterns in residues flexibility occur in both apo and holo forms. In addition, the ΔRMSF analysis of this complex was represented onto cruzain average structure, which agrees with *CP* results (Figure 25C). Note, that loop$_{10-25}$ and loop$_{41-50}$ increase the flexibility in presence of ligand, and some residues belonging to theses loops were also defined as hot spots in section of per-residue energy decomposition of ZINC00352089 complex.

The simulations run with ZINC83627668 complex indicate a uniform increase in overall protein flexibility, favoring distortions in both, loops and well-defined secondary structures (Figure 25B). This phenomenon is quite unusual since generally the compound binding stabilizes proteins, reducing movement. However, this fact could be a characteristic of certain allosteric inhibitors, which can favor transitions away from basal catalytic states of enzyme. Noteworthy, that ZINC83627668 is positive charged at pH employed for MD simulations (pH=5.5, see Table 1), which could induce a perturbation on protein movements and electrostatic balance. However, we observed no significant changes on protein surface electrostatic potential subsequent to this ligand binding (data not shown). The ΔRMSF results also showed appreciable differences between ZINC83627668 and ZINC00352089 complexes fluctuations, suggesting that binding to the same allosteric site does not necessarily induce the same effect on protein structure (Figure 25C and D).

Finally, the RMSD distribution of the three analyzed systems, i.e., the apo form and the two complexes, shows three different, yet partially overlapping populations. Interestingly, a noticeable increase in RMSD values of ZINC83627668 complex was observed (0.5Å, see Figure 25F). On the other hand, the intra-molecular interactions, i. e., hydrogen bonds, salt bridges and van der Waals interactions were calculated for each system along the simulation time. This analysis shows no differences in hydrophobic contacts between the three systems (data not shown). Nevertheless, some dissimilarities in hydrogen bond patterns were observed in site 3 vicinity between apo and holo forms (Figure 25E). As it was expected, the presence of ligands in site 3 led to the disruption of salt bridge forming by LYS17 and GLU35. Specifically, hydrogen bonds between site 3 hot spots, i. e., THR14, LYS17, LEU48, and GLU35, are missing in holo form (Figure 25E). All these breaks of intra-molecular interactions could explain, to a certain extent, the aforementioned changes in cruzain internal dynamics.

4.7. Small changes along PC2 detected for cruzain apo and holo forms

For further comprehension of cuzain correlated dynamics, PCA was performed using the Cα covariance matrix (Figure 26). We can observe that the first 12 eigenvectors represent around 49% of total atomic fluctuations of enzyme, which means that the most of total variance is not represented by only the first few eigenvectors. This is an expected result for quasi-rigid proteins (Figure 26A and B), considering the represented eigenvalues relative to constrained and more localized fluctuation. Interestingly, previous works in which PCA was performed for cruzain and papain-like cysteine proteases systems did not show the statistical information of eigenvalues in terms of proportion of variance (Hoelz,Leal *et al.* 2016;Novinec 2017), preventing to compare with our results.

We consider to represent our results in terms of PC1 and PC2 in cruzain apo and holo forms, since previous PCA performed for the family of papain-like cysteine proteases only considered the collective modes of these PCs to describe most relevant motions in these enzymes (Hoelz,Leal *et al.* 2016;Novinec 2017). The representation of PC subspace as 2D histograms shows the density population of each conformational state of cruzain systems (Figure 26C and D). The projection of the movements of the holo states respect onto the apo PC2, confirms that MD conformers of apo form sampled a broader conformational space compared to the holo ones. Noteworthy, the MD simulation of the apo form is a concatenation of five replicas performed with different cruzain crystal structures. These initial conditions of MD simulation may enhance the sampling of the conformational space. In addition, it is well known that ligand binding tends to stabilize the protein structures, which could decreases the movements of great amplitude in holo forms.

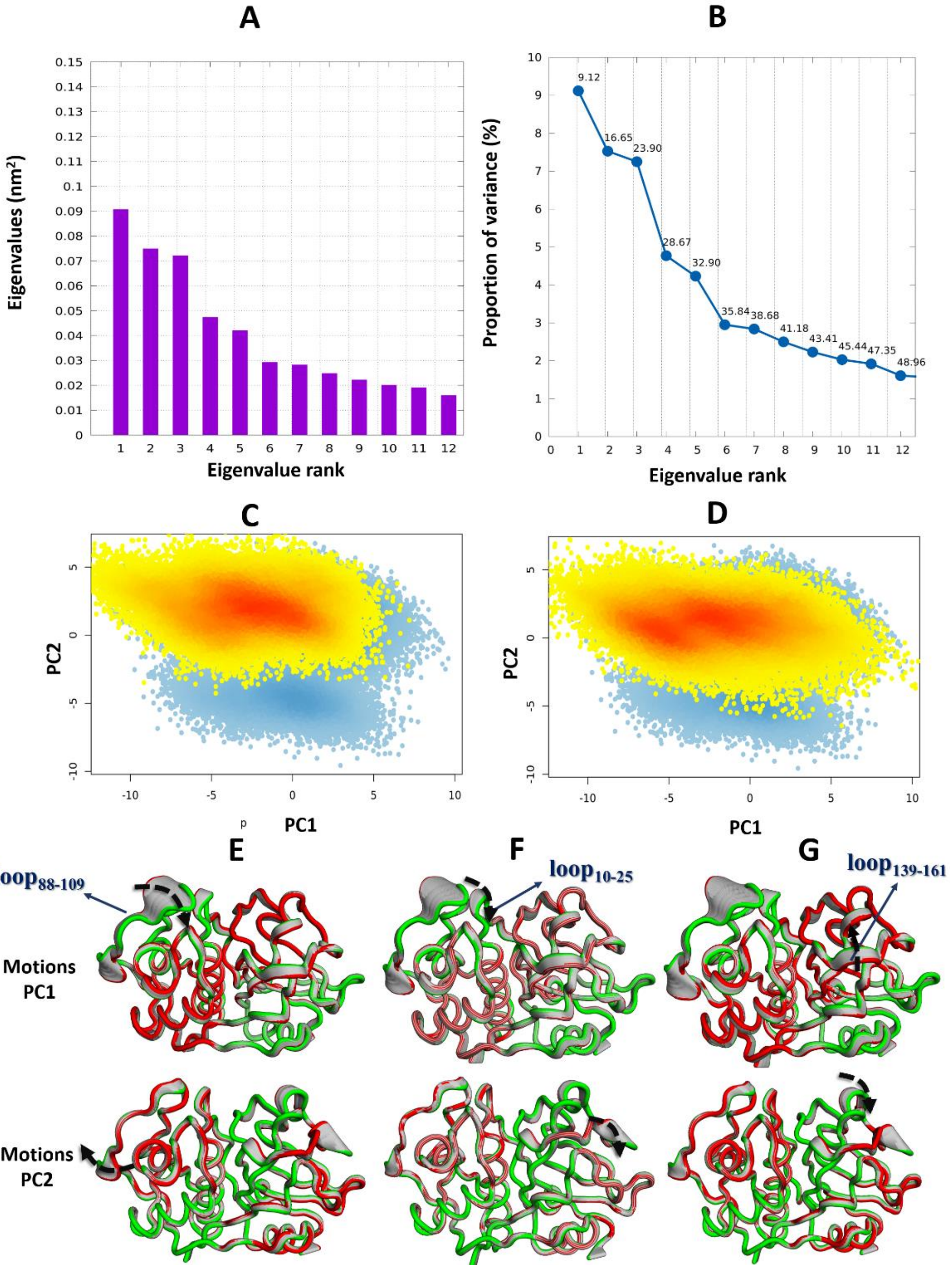


Figure 26. Principal component analysis (PCA) of cruzain apo and holo forms. *(A) Amplitude of first 12 eigenvalues calculated from the covariance matrix of Cα coordinates from MD simulations. (B) The graph of each eigenvalue magnitude expressed as the percentage of the total variance captured by the corresponding eigenvector. Labels beside each point indicate the cumulative sum of the total variance of all preceding eigenvectors. The projection of trajectory was made onto the principal planes defined by the first two principal components. The holo form (orange) was projected onto the eigenvectors of the apo form (blue) for (C) cruzain-ZINC00352089 and (D) cruzain-ZINC83627668 complexes. A color gradient was employed to represent the density of structures in each region of phase space. The projections of the MD motions were represented along the first two eigenvector for (E) the apo form, (F) cruzain-ZINC00352089 and (G) cruzain-ZINC83627668 systems. The black arrows show the direction of collective motions in the sense from red to green and the principal loops are labeled.*

The structures depicted in Figure 26E, F and G represent the extreme snapshots (extended structures) along the direction of collective motions (from red to green direction) described by eigenvector 1 and eigenvector 2. In both systems (apo and holos), the two PC modes were associated mainly with loop displacements, in accordance with previous results of per-residue RMSF values (sections 4.2 and 4.6). Along the PC1 vector, the Cα of loop$_{88\text{-}109}$ undergo an open-close-like motion while the cruzain active site remains relatively in a fixed position. The PC2 motion, where more dissimilarities between apo and holo were expected according to Figure 26E, F and G, only subtle differences in amplitude of the movement of 143-146 region were observed. Noteworthy, papain-like cysteine proteases contain three/four disulfide bonds which stiffen their structures. Also, the eigenvector 1 and 2 only represent two of the most dominant degrees of freedom in the configurational space, so the projected structures do not necessarily correspond to physical structures sampled during the MD simulations.

Previous works of PCA performed with papain-like cysteine proteases also denoted that the essential dynamics of these proteins lies in loop regions (Hoelz,Leal *et al.* 2016;Novinec 2017). In addition, all studies highlighted the pivotal role of loop$_{88\text{-}109}$ in cruzain motions. Moreover. Hoelz *et al.* (Hoelz,Leal *et al.* 2016) reported that the identified movements for the apo form display an opening of the structure that exposes its active site (open conformation). Conversely, our results do not reproduce this behavior, which may be influenced by the conditions and structures employed in MD simulations.

4.8. Pairwise distance distributions identify differences in key residues of S2 and S3 subsites between the apo and holo forms

The action of allosteric inhibitors ultimately leads to structural changes that alter the enzyme activity. The results in the previous sections indicate that only minor conformational changes occurred in loop regions of cruzain structure. However, they might be sufficient to modify the enzymatic activity; since, as demonstrated elsewhere, the allosteric inhibition of many enzymes involves just subtle structural changes (Tsai,del Sol *et al.* 2008;Novinec 2017).

A comparison of the distribution of interatomic distances in cruzain shows some variability in the S2 and S3 regions formed by loop$_{56\text{-}67}$ and loop$_{197\text{-}208}$. In particular, distances between residue pairs shown in Figure 27 were identified as those displaying the most appreciable differences in the active site (rest of the data not shown). As can be observed, the distance distributions of the apo and holo (enzyme+allosteric inhibitor)

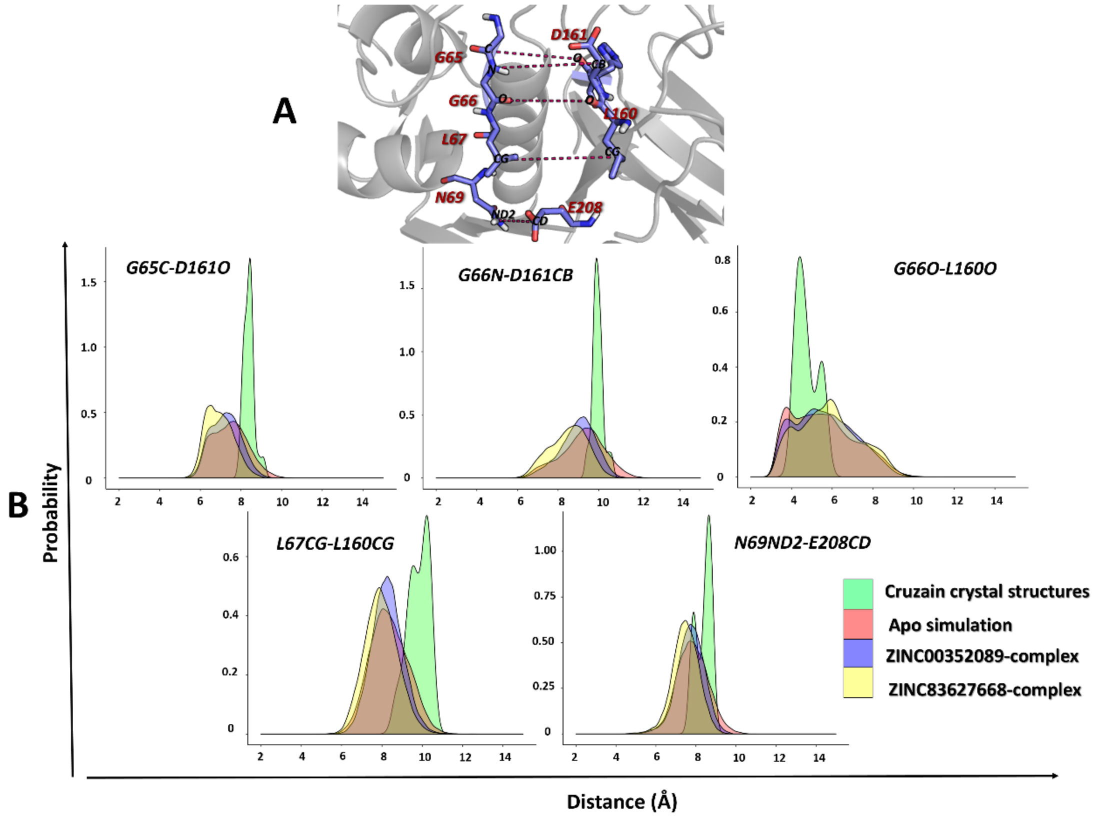


__Figure 27. Distributions of pairwise interatomic distance of cruzain binding site.__ (A) S2-S3 region of cruzain with the interatomic distances of its principal residues. Selected residues are labeled and their side-chains are depicted as stick. (B) Distance distributions obtained from the MD simulations of three analyzed systems and from cruzain crystal structures. Graphs are labeled with the atomic pair analyzed in each case.

forms have, in most cases, their respective maxima shifted towards lower values when compared to those of the crystal structures of cruzain bound to the orthosteric inhibitors. This suggests that the S3 and S2 subsites is slightly narrower when there is no ligand bound to it. Therefore, the active site is likely to undergo a small structural reorganization to adapt to anchor other molecules. On the other hand, the differences in the distance distributions of the apo and both holo forms highlight the occurrence of structural changes resulting from the binding of ligands to the putative allosteric site. The link between such dynamical differences in the active site and the inhibition of the protease remains elusive. Note, however, that a previous work (Novinec 2017) has emphasized the occurrence of different residue-residue distance distributions across the active site of HCatK in the apo and allosterically inhibited forms. Therefore, this phenomenon seems to be a hallmark of allosteric inhibition, and might be of major relevance for systems lacking large conformational changes during allosteric modulation.

4.9. Correlation network analysis reveals state-specific differences in residue couplings

Comparison of cross-correlation (*CC*) and generalized correlation (*GC*) approaches shows that inter-residue correlation is strongest when using *GC* (Figure 28). Both methods correctly identify highly coupled motions in residues belonging to substrate specificity subsites S2 and S3 (violet boxes). In addition, residues of loop$_{84-109}$ (belonging to site 1) and residues of helix 4 (conforming site 3) possess correlation values above 0.6.

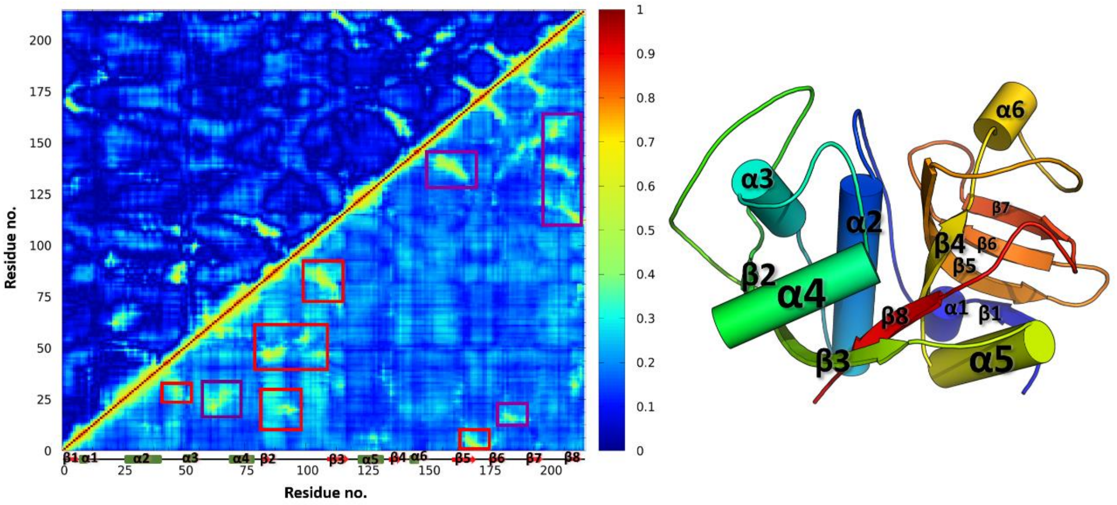


Figure 28. Comparison of the standard cross-correlation with generalize correlation coefficient values. *Upper triangle corresponds to standard cross-correlation and lower triangle the generalize correlation. The framed values are related to strong correlations where the violet squares point to the residues belonging to the catalytic core and to specificity subsites of cruzain. On the right side, the secondary structures of cruzain are numbered from N to C-terminal. This nomenclature is also used in a linear sequence of cruzain at the bottom of the correlation matrix.*

In parallel, the results obtained for apo and holo enzyme reveal an increase of correlations, principally in cruzain-ZINC83627668 complex, triggered by ligand binding. This is demonstrated by the prevalence of positive values in difference correlation matrix, i.e., holo minus apo correlations (pink spots) (Figure 29). In the case of cruzain-ZINC00352089 complex, a visible increment in correlation values is observed for residues located in positions 14-19, 45-50, 90-100 (Figure 29A), while in cruzain-ZINC83627668 system, almost all values increased (Figure 29B). Besides, a decrease of correlation values (green points) of helix 2 (CYS25 localization) is observed in both systems (Figure 29). Changes in coupled motions corresponding to interface residues of holo form are associated with the establishment of a "linkage" between loop$_{10-25}$ and loop$_{84-109}$ mediated by ligand.

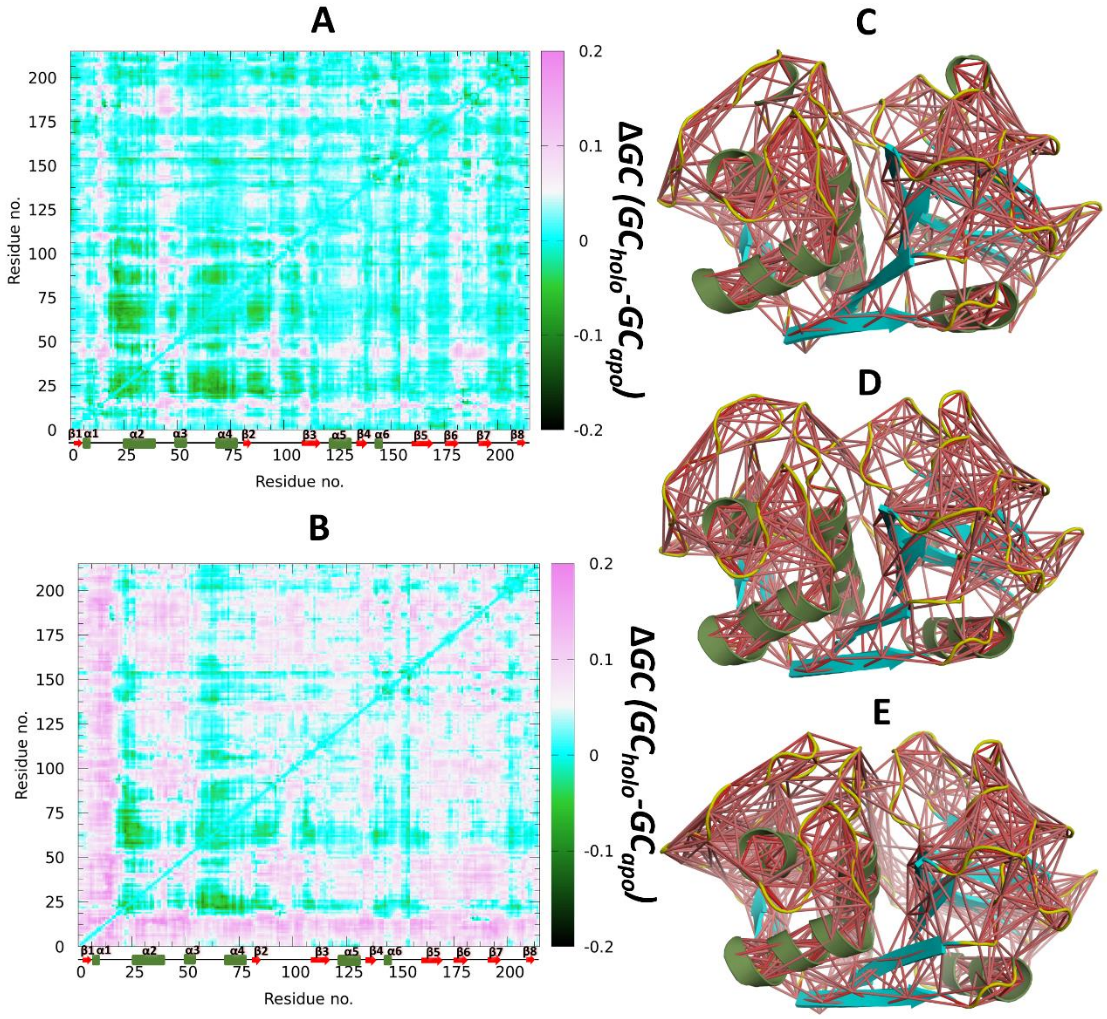


***Figure 29. Comparison of generalized correlation coefficients between apo and holo forms of cruzain**. Matrices corresponding to differences of pairwise GC values are represented for (A) cruzain-ZINC00352089 and (B) cruzain-ZINC83627668 complexes. Correlation networks derived from (C) apo simulations, (D) cruzain-ZINC00352089 and (E) cruzain-ZINC83627668 complexes. The edge color intensity is proportional to correlation values.*

The correlation network analysis further dissects residue couplings and reveals residue communities having the potential to facilitate long-range allosteric signal propagation. In addition, an analysis of intercommunity coupling strength can reveal the state specific coupling paths and overall dynamical packing in cruzain. The structure of papain-like proteases family can be clearly partitioned into two main domains, and the analysis of time-averaged weighted networks adequately captured both of them in all studied systems (Figure 30). In network-based representation of cruzain apo form, its binding site is allocated in two different communities, i. e., 9 and 5, while the two domains are linked by communities 7-9, 7-8 and 2-3. This highlights the indirect coupling of functional domains through the catalytic site. Analysis of intercommunity coupling

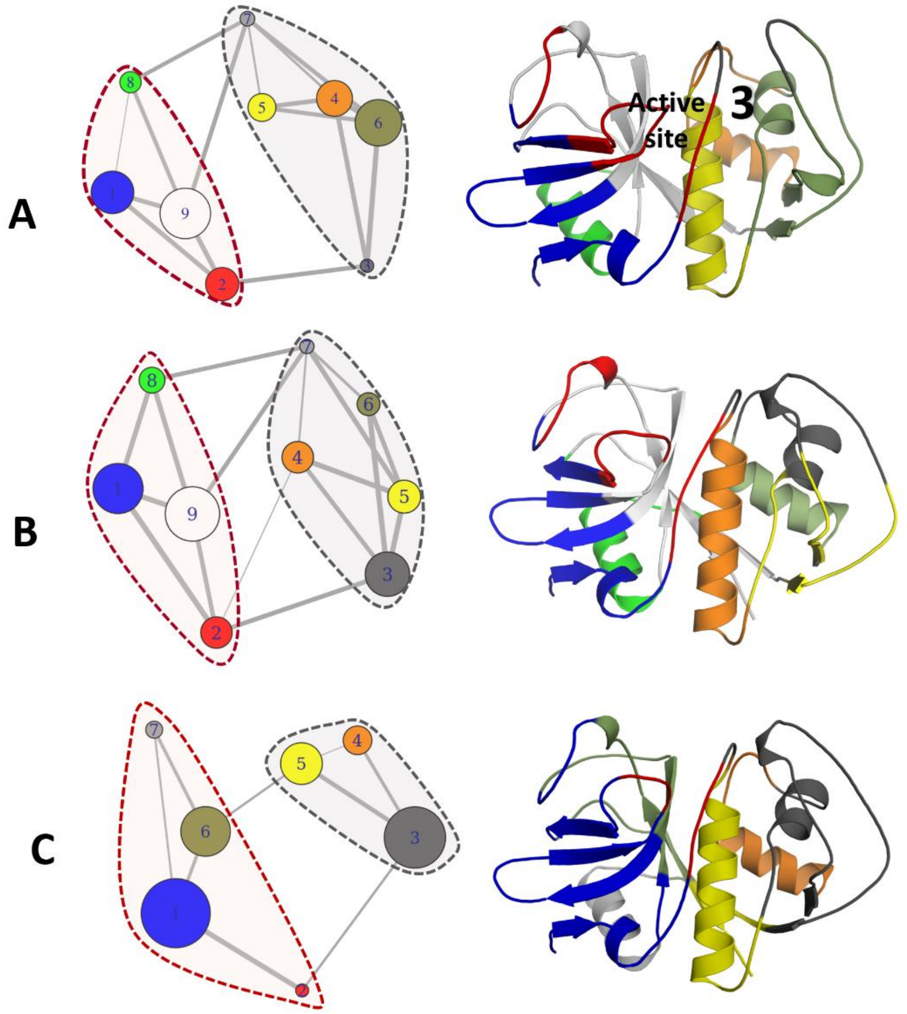


Figure 30. ***Community network analysis.*** *Community structure of (A) cruzain apo form, (B) cruzain-ZINC00352089 and (C) cruzain-ZINC83627668 systems. Dashed lines delimit the cruzain structural domains and edge thickness scales are proportional to the number of shortest path passing through those junctions. Communities' size is proportional to number of their composite residues.*

strengths reveals state specific allosteric coupling between site 3 and active site through the link between community 6 and 5, and by connection of community 2 and 9 (Figure 30A).

The comparison of the community structure in isolated cruzain and complexes reveals remarkable conformational changes induced by effector binding (Figure 30B, C and Table 3), highlighting a rearrangement of community numbers and its residue members in holo systems. The cruzain-ZINC00352089 complex displays a change in community 3 size and a new weak coupling between the main domains appears (edge between community 4-2 and 4-3). However, the strength of direct interactions remain unchanged respect to the apo form, except in the interaction between communities 1 and

8, which enhance its correlation. Note that communities reordering take place in alpha-helix domain (delimited by gray dashed lines in Figure 30B), specifically in loop$_{41-50}$ and loop$_{84-109}$, which are located within site 3 and site 1, respectively (Figure 30B).

Conversely, in the cruzain-ZINC83627668 complex a whole rearrangement in communities is observed (in both protein domains) (see communities delimited by gray and red dashed lines in Figure 30C and see Table 3). Here, the domain containing the helix 2 (delimited by gray dashed lines) lost two communities, inducing a “packing” within its structure. On the other hand, ZINC83627668 binding induces changes in domain of beta-sheets (delimited by red dashed lines), with reassignment of a few residues from communities 2, 8 and 9 in a newly formed community 1. The decrease of community numbers suggests that compound ZINC83627668 triggers an overall structural reorganization of cruzain. The strength of inter and intra coupling decreases considerably in this system, which could affect the overall enzyme activity. Indeed, the appearance of a new correlation between two domains through communities containing residues of active site is observed, indicates a perturbation in the correlations of these residues (edge between community 6 and 5 in Figure 30C).

Table 3. Distribution of cruzain residues within optimal community structures of apo and ligand bound complexes.

Community ID	Apo	ZINC00352089 complex	ZINC83627668 complex
1	1-13, 151-152, 167-179, 195-198	1-13, 150-153, 167-179, 194-198	1-14, 133-135, 146-152, 165-182, 188-198
2	14-18, 146-150, 180-194	14-19, 145-149, 183-193	15-19, 183-187
3	19-22, 92-97	20-23, 51-66, 92-102	20-22, 47-59, 81-111
4	23-24, 57-81	24-45	23, 60-80
5	25-45	46-50, 83-91, 103-111	24-46, 112-115, 211-215
6	46-56, 82-91, 98-111	67-82	116-118, 120, 136-145, 153-164, 199-210
7	112-116, 210-215	112-116, 211-215	119, 121-132
8	117-132	117-132, 199-200	-
9	133-145, 153-166, 199-209	133-144, 154-166, 180-182, 201-210	-

4.10. Network path analysis reveals couplings between cruzain orthosteric site and allosteric site 3

The residue centrality based on average betweenness can characterize and differentiate highly connected residues that mediate stable interaction networks and allosteric communications in protein structures. This parameter can be presented as profiles, being a guiding indicator for the identification of functional residues critical for allosteric regulation. The centrality analysis highlights remarkable differences in the way information passes through free cruzain versus enzyme bound to allosteric modulators (Figure 31). These results show changes in the centrality degree upon ligand association, indicating an alteration in dynamic couplings and a reduction in pathway diversity in the shortest pathways that connect all residues.

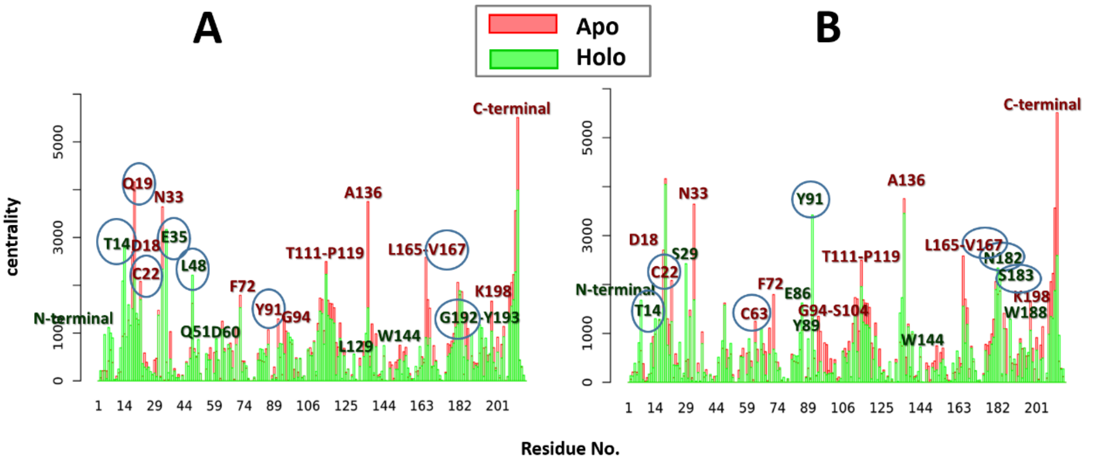


Figure 31. Residue centrality in cruzain analyzed systems. *Comparison of node centrality values for (D) cruzain-ZINC00352089 and (E) cruzain-ZINC83627668 are represented. Residues with the highest values of centrality are labeled in red and green for apo and holo systems, respectively. Conserved residues in papain-like cysteine proteases are marked with blue circles.*

In the apo form, some residues with high-density network connections, i.e., ASP18, GLN19, ASN33, HIS115, ALA136, LEU165 and residues belonging to the C-terminal region were detected (Figure 31). Note that all these residues lie at the interface of the two domains of cruzain, which points out that the high level of signal pathways that cross over them may be connecting these two lobes to favor enzyme function. It is noteworthy the high centrality values of GLN19, which is a key residue in catalysis of papain-like proteases.

In cruzain-ZINC00352089 complex new highly connected residues arise, i. e., THR14, GLU35, LEU48 and N-terminal region (Figure 31A), where the three former ones are laid in the site 3 region. These results point out that ligand presence increases the number of shortest pathways crossing over site 3, making this groove a potential region for control of enzyme function under these conditions. In the case of cruzain-ZINC83627668 complex, residues SER29, GLU86, TYR91, ASN182 and SER183 increase their centrality (Figure 31B). Note that also GLU86, TYR91, ASN182 and SER183 are part of site 3. Interestingly, all residues mentioned above are conserved across the papain-like cysteine proteases (Durrant,Keranen *et al.* 2010). This indicates that ligands designed against site 3 generate an emergent connection involving common residues of this protease family, which would be unfavorable for enzyme catalytic function.

The optimal and suboptimal pathways were calculated between CYS25 and each hot spot previously identified for site 3. However, only the analysis performed for CYS25-THR14 pair showed that ligand binding strengthens the correlation between site 3 and catalytic core (rest of the data not shown). This is consistence with centrality analysis where the THR14 was one of the main residue displaying significant changes in its connections as a consequence of ligand binding (Figure 31).

Examination of path lengths distribution indicates that the shortest pathways of cruzain-ZINC00352089 complex and apo enzyme do not differ substantially (apo, 3.12; holo, 2.9) (Figure 32C). Nevertheless, the distribution derived from this system trajectory is slightly deviated toward shorter path lengths, indicating that motions of the residues connecting the allosteric and catalytic sites are more tightly correlated when ZINC00352089 is bound. The determination of critical residues for allosteric transmission (residue with highest degeneracy values), reveals different paths between the free enzyme and cruzain-ZIN00352089 complex (Figure 32E, G). Here a considerable decrease in node degeneracy values and the increment in the number of residues involved in shortest paths are observed (Figure 32A). Therefore, this analysis identifies two main routes for allosteric communication where one goes through helix 2 and the other by the loop$_{10-25}$ (Figure 32E, G). Consequently, when taking the pathways altogether, we can conclude that the allosteric consequences of ZINC00352089 binding are relevant.

On the other hand, the shorter paths between site 3 and catalytic site transverse the same nodes in cruzain-ZINC83627668 and apo form (Figure 32B, F). Interestingly,

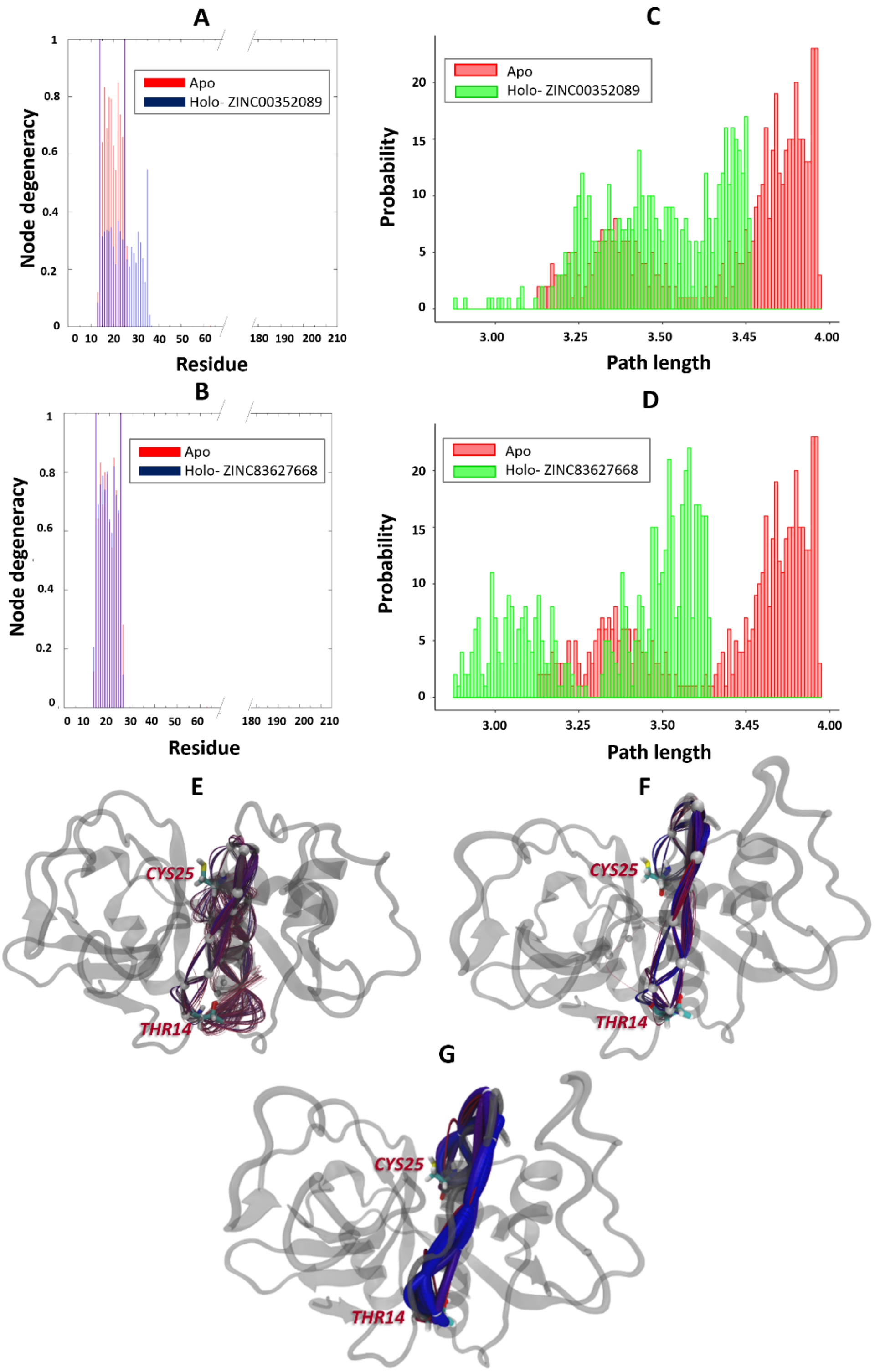


*Figure 32. **Structural representation and distribution of suboptimal pathways.** Node degeneracy in signaling pathways reported for (A) ZINC00352089 and (B) ZINC83627668 systems. Blue color indicates the holo state and red indicates the apo state. A histogram of the 500 path lengths associated with the apo and holo trajectories is shown for both systems: (C) ZINC00352089 and (D) ZINC83627668, comparing with apo distribution. The 500 shortest paths between THR14 and CYS25, derived from (E) the cruzain-ZINC00352089 trajectory, (F) the cruzain-ZINC83627668 trajectory and (G) apo trajectory are shown as blue splines.*

the signal traverses directly to the catalytic core through the residues conforming loop$_{10-}$

$_{25}$. Even though the two systems share the same path, a strongest correlation between site 3 and the catalytic site in the holo form is detected. This becomes apparent by observing that the shortest paths distribution in cruzain-ZINC83627668 system shift toward the direction of lowest values (apo, 3.12; holo, 2.88) (Figure 32D). The previous shift is likely to arise from a more coherent signal in the holo simulation, indicating a possible decrease in the entropy along the pathways due to ZINC83627668 binding. Furthermore, an increment in shortest paths number is observed in cruzain-ZINC8627668 with respect to cruzain-ZINC00352089 complex, and the probability values of such paths are bigger in the former system than in the latter (Figure 32C, D). All these lines of evidence point out a dynamical tightening of the active site with site 3 in presence of ZINC83627668, which suggests, more potent allosteric influences of this compound in comparison with the other one.

Previous studies in evolutionary conservation have shown that most residues of loop$_{10-25}$ and helix 2 are partially or strongly conserved across papain-like proteases. (Durrant,Keranen *et al.* 2010). The changes in dynamical tightening indicate a possible regulation of key allosteric residues of this protein family, but also highlight the possible specificity of compounds which target this protein region. All these results evidence an effective signal propagation which, in turn, may enable the allosteric modulation.

5. Conclusions

Our results constitute a strategy for designing novel allosteric inhibitors of cruzain, the major protease of *T. cruzi*. First, the dynamical survey of previously identified cavities evidenced the presence of one stable pocket with a functional gate (site 3) and one transient pocket (site 1). This characterization of site 1 and site 3 was not be feasible with a simple analysis or superposition of static crystallographic structures of cruzain. Our study on these allosteric sites, which are druggable by small molecules and can modulate the enzymatic activity (Novinec,Lenarcic *et al.* 2014), provides a solid basis for further drug discovery. Secondly, the combination of Autodock Vina score and MM-GBSA free energy calculations allowed to us to propose compounds ZINC00352089 and ZINC83627668 as scaffolds of cruzain allosteric modulators. In addition, we verified that the aliphatic chains of the residues of site 3 mediated the main protein-ligand interactions within the groove. These aspects can be important implications for targeted optimization of lead compounds or for research purposes.

Finally, the allosteric pathways which link site 3 with active site were elucidated by combining MD simulations with correlation of protein motions based in network theory. In this context, no major effects on active site structure were observed due to compound binding (modification of distance and angles between catalytic residues), which indicates that allosteric regulation in cruzain is mediated via alterations of its dynamical properties similarly to HCatK allosteric inhibition (Novinec,Rebernik *et al.* 2016). Mapping communication pathways between site 3 and the catalytic core shows that ZINC83627668 triggers the propagation of dynamics, creating shorter pathways, and stronger correlations. According to our results compound ZINC83627668 disturbs cruzain structure more than ZINC00352089, pointing out that structural dissimilarities between ligands may defined different communication routes, even though they are linked to same allosteric site. In summary, the disruption of this specific network communication could represent a rational approach for designing drugs with improved potency and selectivity against cruzain enzymatic function.

6. References

Almeida, P. C., *et al.* (2001). "Cathepsin B activity regulation. Heparin-like glycosaminogylcans protect human cathepsin B from alkaline pH-induced inactivation." J Biol Chem **276**(2): 944-951.

Almeida, P. C., *et al.* (1999). "Cysteine proteinase activity regulation. A possible role of heparin and heparin-like glycosaminoglycans." J Biol Chem **274**(43): 30433-30438.

Amadei, A., *et al.* (1993). "Essential dynamics of proteins." Proteins **17**(4): 412-425.

Arafet, K., *et al.* (2017). "Computational Study of the Catalytic Mechanism of the Cruzain Cysteine Protease." ACS Catalysis **7**(2): 1207-1215.

Bashford, D. and D. A. Case (2000). "Generalized born models of macromolecular solvation effects." Annu Rev Phys Chem **51**: 129-152.

Bayly, C. I., *et al.* (1993). "A well-behaved electrostatic potential based method using charge restrains for deriving atomic charges: The RESP model." J Phys Chem **97**: 10269-10280.

Beaulieu, C., *et al.* (2010). "Identification of potent and reversible cruzipain inhibitors for the treatment of Chagas disease." Bioorg Med Chem Lett **20**(24): 7444-7449.

Berasain, P., *et al.* (2003). "Specific cleavage sites on human IgG subclasses by cruzipain, the major cysteine proteinase from *Trypanosoma cruzi*." Mol Biochem Parasitol **130**(1): 23-29.

Berman, H. M., *et al.* (2000). "The protein data bank." Nucleic Acids Res **28**(1): 235-242.

Besler, B. H., *et al.* (1990). "Atomic charges derived from semiempirical methods " J Compu Chem **11**: 431-439.

Branquinha, M. H., *et al.* (2015). "Cruzipain: An Update on its Potential as Chemotherapy Target against the Human Pathogen *Trypanosoma cruzi*." Curr Med Chem **22**(18): 2225-2235.

Brooijmans, N. and I. D. Kuntz (2003). "Molecular recognition and docking algorithms." Annu Rev Biophys Biomol Struct **32**: 335-373.

Case, D. A., *et al.* (2014). Amber 14.

Collier, G. and V. Ortiz (2013). "Emerging computational approaches for the study of protein allostery." Arch Biochem Biophys **538**(1): 6-15.

Cooper, A. and D. T. Dryden (1984). "Allostery without conformational change. A plausible model." Eur Biophys J **11**(2): 103-109.

Costa, M. G., *et al.* (2010). "How does heparin prevent the pH inactivation of cathepsin B? Allosteric mechanism elucidated by docking and molecular dynamics." BMC Genomics **11 Suppl 5**: S5.

Costa, T. F., *et al.* (2012). "Substrate inhibition and allosteric regulation by heparan sulfate of *Trypanosoma* brucei cathepsin L." Biochim Biophys Acta **1824**(3): 493-501.

Coura, J. R. and J. Borges-Pereira (2010). "Chagas disease: 100 years after its discovery. A systemic review." Acta Trop **115**(1-2): 5-13.

Coura, J. R., *et al.* (2014). "Ecoepidemiology, short history and control of Chagas disease in the endemic countries and the new challenge for non-endemic countries." Mem Inst Oswaldo Cruz **109**(7): 856-862.

Chemical Computing Group ULC (2017). "Molecular Operating Environment (MOE)." 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.

Chen, Y. C. (2015). "Beware of docking!" Trends Pharmacol Sci **36**(2): 78-95.

Christ, C. D. and T. Fox (2014). "Accuracy assessment and automation of free energy calculations for drug design." J Chem Inf Model **54**(1): 108-120.

D.A. Case, D. S. C., T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (2017). "AMBER 2017." University of California **San Francisco**.

da Silva, E. B., *et al.* (2016). Trypanosomal Cysteine Peptidases: Target Validation and Drug Design Strategies. Comprehensive Analysis of Parasite Biology: From Metabolism to Drug Discovery, Wiley-VCH Verlag GmbH & Co. KGaA**:** 121-145.

Dalafave, D. S. and R. E. Dalafave "Computational Design of Allosteric Inhibitors of AKT and SGK Kinases." Biophys J **108**(2): 320a.

Darden, T., *et al.* (1993). "Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems." J Chem Phys **98**(12): 10089-10092.

de Ruiter, A. and C. Oostenbrink (2011). "Free energy calculations of protein-ligand interactions." Curr Opin Chem Biol **15**(4): 547-552.

De Vivo, M., *et al.* (2016). "Role of Molecular Dynamics and Related Methods in Drug Discovery." J Med Chem **59**(9): 4035-4061.

DeLano, W. (2004). "Use of PyMOL as a communications tool for molecular science." Abst Pap Am Chem Soc **228**: U313-U314.

Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs." Numer. Math. **1**(1): 269-271.

Dokholyan, N. V. (2016). "Controlling Allosteric Networks in Proteins." Chem Rev **116**(11): 6463-6487.

Dolinsky, T. J.*, et al.* (2004). "PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations." Nucleic Acids Res **32**(suppl 2): W665-W667.

Doyle, P. S.*, et al.* (2007). "A cysteine protease inhibitor cures Chagas' disease in an immunodeficient-mouse model of infection." Antimicrob Agents Chemother **51**(11): 3932-3939.

Doyle, P. S.*, et al.* (2011). "The *Trypanosoma cruzi* protease cruzain mediates immune evasion." PLoS Pathog **7**(9): e1002139.

Drag, M. and G. S. Salvesen (2010). "Emerging principles in protease-based drug discovery." Nat Rev Drug Discov **9**(9): 690-701.

Durrant, J. D.*, et al.* (2011). "POVME: an algorithm for measuring binding-pocket volumes." J Mol Graph Model **29**(5): 773-776.

Durrant, J. D.*, et al.* (2010). "Computational identification of uncharacterized cruzain binding sites." PLoS Negl Trop Dis **4**(5): e676.

Duschak, V. G.*, et al.* (2001). "Enzymatic activity, protein expression, and gene sequence of cruzipain in virulent and attenuated *Trypanosoma cruzi* strains." J Parasitol **87**(5): 1016-1022.

Ferrao, P. M.*, et al.* (2015). "Cruzipain Activates Latent TGF-beta from Host Cells during *T. cruzi* Invasion." PLoS One **10**(5): e0124832.

Ferreira, L. G. and A. D. Andricopulo (2017). "Targeting cysteine proteases in trypanosomatid disease drug discovery." Pharmacol Ther.

Ferreira, L. G.*, et al.* (2015). "Molecular Docking and Structure-Based Drug Design Strategies." Molecules **20**(7): 13384-13421.

Ferreira, R. S.*, et al.* (2010). "Complementarity Between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors." J Med Chem **53**(13): 4891-4905.

Frisch, M. J.*, et al.* (2009). "Gaussian 09." Gaussian Inc **Wallingford CT**.

Gentry, P. R.*, et al.* (2015). "Novel Allosteric Modulators of G Protein-coupled Receptors." J Biol Chem **290**(32): 19478-19488.

Gillmor, S. A., *et al.* (1997). "Structural determinants of specificity in the cysteine protease cruzain." Protein Sci **6**(8): 1603-1611.

Girvan, M. and M. E. Newman (2002). "Community structure in social and biological networks." Proc Natl Acad Sci U S A **99**(12): 7821-7826.

Gohlke, H., *et al.* (2003). "Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes." J Mol Biol **330**(4): 891-913.

Gohlke, H. and G. Klebe (2002). "Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors." Angew Chem Int Ed Engl **41**(15): 2644-2676.

Gourbiere, S., *et al.* (2012). "Genetics and evolution of triatomines: from phylogeny to vector control." Heredity (Edinb) **108**(3): 190-202.

Groebe, D. R. (2006). "Screening for positive allosteric modulators of biological targets." Drug Discov Today **11**(13-14): 632-639.

Groebe, D. R. (2009). "In search of negative allosteric modulators of biological targets." Drug Discov Today **14**(1-2): 41-49.

Grover, A. K. (2013). "Use of allosteric targets in the discovery of safer drugs." Med Princ Pract **22**(5): 418-426.

Gunasekaran, K., *et al.* (2004). "Is allostery an intrinsic property of all dynamic proteins?" Proteins **57**(3): 433-443.

Hauske, P., *et al.* (2008). "Allosteric regulation of proteases." Chembiochem **9**(18): 2920-2928.

Hernandez-Rodriguez, M., *et al.* (2016). "Current tools and methods in Molecular Dynamics (MD) simulations for drug design." Curr Med Chem.

Hertig, S., *et al.* (2016). "Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations." PLoS Comput Biol **12**(6): e1004746.

Hilser, V. J., *et al.* (2012). "Structural and energetic basis of allostery." Annu Rev Biophys **41**: 585-609.

Hoelz, L. V., *et al.* (2016). "Molecular dynamics simulations of the free and inhibitor-bound cruzain systems in aqueous solvent: insights on the inhibition mechanism in acidic pH." J Biomol Struct Dyn **34**(9): 1969-1978.

Homeyer, N. and H. Gohlke (2012). "Free energy calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area method." Mol Inform **31**(2): 114-122.

Hotez, P. and D. A. P. Bundy (2017). "The PLOS Neglected Tropical Diseases decade." PLoS Negl Trop Dis **11**(4): e0005479.

Hou, T., *et al.* (2011). "Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations." J Chem Inf Model **51**(1): 69-82.

Huey, R. and G. M. Morris (2008). "Using AutoDock 4 with AutoDocktools: a tutorial." The Scripps Research Institute, USA: 54-56.

Humphris, E. L. and T. Kortemme (2008). "Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design." Structure **16**(12): 1777-1788.

Hunenberger, P. H., *et al.* (1995). "Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations." J Mol Biol **252**(4): 492-503.

Jayaram, B., *et al.* (1998). "Solvation Free Energy of Biomacromolecules: Parameters for a Modified Generalized Born Model Consistent with the AMBER Force Field." The Journal of Physical Chemistry B **102**(47): 9571-9576.

Jilkova, A., *et al.* (2014). "Activation route of the *Schistosoma mansoni* cathepsin B1 drug target: structural map with a glycosaminoglycan switch." Structure **22**(12): 1786-1798.

Jorgensen, W. L. and C. Jenson (1998). "Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density." J Comput Chem **19**(10): 1179-1186.

Judice, W. A., *et al.* (2013). "Heparin modulates the endopeptidase activity of *Leishmania mexicana* cysteine protease cathepsin L-Like rCPB2.8." PLoS One **8**(11): e80153.

Kenakin, T. P. (2010). "Ligand detection in the allosteric world." J Biomol Screen **15**(2): 119-130.

Kleinjung, J. and F. Fraternali (2014). "Design and application of implicit solvent models in biomolecular simulations." Curr Opin Struct Biol **25**: 126-134.

Korb, O., *et al.* (2012). "Potential and limitations of ensemble docking." J Chem Inf Model **52**(5): 1262-1274.

Koshland, D. E., Jr., *et al.* (1966). "Comparison of experimental binding data and theoretical models in proteins containing subunits." Biochemistry **5**(1): 365-385.

Kumar, S., *et al.* (1999). "Folding funnels and conformational transitions via hinge-bending motions." Cell Biochem Biophys **31**(2): 141-164.

Lange, O. F. and H. Grubmuller (2006). "Generalized correlation for biomolecular dynamics." Proteins **62**(4): 1053-1061.

Latorre, A., *et al.* (2016). "Dipeptidyl Nitroalkenes as Potent Reversible Inhibitors of Cysteine Proteases Rhodesain and Cruzain." ACS Med Chem Lett **7**(12): 1073-1076.

Li, Z., *et al.* (2004). "Regulation of collagenase activities of human cathepsins by glycosaminoglycans." J Biol Chem **279**(7): 5470-5479.

Lima, A. P., *et al.* (2002). "Heparan sulfate modulates kinin release by *Trypanosoma cruzi* through the activity of cruzipain." J Biol Chem **277**(8): 5875-5881.

Lionta, E., *et al.* (2014). "Structure-based virtual screening for drug discovery: principles, applications and recent advances." Curr Top Med Chem **14**(16): 1923-1938.

Liu, J. and R. Nussinov (2016). "Allostery: An Overview of Its History, Concepts, Methods, and Applications." PLoS Comput Biol **12**(6): e1004966.

Lu, S., *et al.* (2014). "Recent computational advances in the identification of allosteric sites in proteins." Drug Discov Today **19**(10): 1595-1600.

Lu, S., *et al.* (2014). "Harnessing allostery: a novel approach to drug discovery." Med Res Rev **34**(6): 1242-1285.

Lugowska, I., *et al.* (2015). "Trametinib: a MEK inhibitor for management of metastatic melanoma." Onco Targets Ther **8**: 2251-2259.

Ma, B., *et al.* (1999). "Folding funnels and binding mechanisms." Protein Eng **12**(9): 713-720.

Ma, B. and R. Nussinov (2014). "Druggable orthosteric and allosteric hot spots to target protein-protein interactions." Curr Pharm Des **20**(8): 1293-1301.

Mackey, T. K., *et al.* (2014). "Emerging and reemerging neglected tropical diseases: a review of key characteristics, risk factors, and the policy and innovation environment." Clin Microbiol Rev **27**(4): 949-979.

Maier, J. A., *et al.* (2015). "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB." J Chem Theory Comput **11**(8): 3696-3713.

Mark W. Robinson PhD, J. P. D. P. (2011). Cysteine Proteases of Pathogenic Organisms, Springer US.

Marques, A. F., *et al.* (2013). "Falcipain-2 inhibition by suramin and suramin analogues." Bioorg Med Chem **21**(13): 3667-3673.

Marques, A. F., *et al.* (2015). "Allosteric regulation of the *Plasmodium falciparum* cysteine protease falcipain-2 by heme." Arch Biochem Biophys **573**: 92-99.

Martinez-Mayorga, K., *et al.* (2015). "Cruzain inhibitors: efforts made, current leads and a structural outlook of new hits." Drug Discov Today **20**(7): 890-898.

May, L. T., *et al.* (2007). "Allosteric modulation of G protein-coupled receptors." Annu Rev Pharmacol Toxicol **47**: 1-51.

McGrath, M. E., *et al.* (1995). "The crystal structure of cruzain: a therapeutic target for Chagas' disease." J Mol Biol **247**(2): 251-259.

Meng, H., *et al.* (2016). "Discovery of Novel 15-Lipoxygenase Activators To Shift the Human Arachidonic Acid Metabolic Network toward Inflammation Resolution." J Med Chem **59**(9): 4202-4209.

Miller, B. R., *et al.* (2012). "MMPBSA.py: An Efficient Program for End-State Free Energy Calculations." J Chem Theory Comput **8**(9): 3314-3321.

Monod, J., *et al.* (1965). "On the Nature of Allosteric Transitions: A Plausible Model." J Mol Biol **12**: 88-118.

Morra, G., *et al.* (2012). "Corresponding functional dynamics across the Hsp90 Chaperone family: insights from a multiscale analysis of MD simulations." PLoS Comput Biol **8**(3): e1002433.

Mott, B. T., *et al.* (2010). "Identification and optimization of inhibitors of Trypanosomal cysteine proteases: cruzain, rhodesain, and TbCatB." J Med Chem **53**(1): 52-60.

Muchmore, S. W. and P. J. Hajduk (2003). "Crystallography, NMR and virtual screening: integrated tools for drug discovery." Curr Opin Drug Discov Devel **6**(4): 544-549.

Ndao, M., *et al.* (2014). "Reversible cysteine protease inhibitors show promise for a Chagas disease cure." Antimicrob Agents Chemother **58**(2): 1167-1178.

Newman, M. E. (2006). "Modularity and community structure in networks." Proc Natl Acad Sci U S A **103**(23): 8577-8582.

Noireau, F., *et al.* (2009). "*Trypanosoma cruzi*: adaptation to its vectors and its hosts." Vet Res **40**(2): 26.

Novinec, M. (2017). "Computational investigation of conformational variability and allostery in cathepsin K and other related peptidases." PLoS One **12**(8): e0182387.

Novinec, M., *et al.* (2014). "A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods." Nat Commun **5**: 3287.

Novinec, M., *et al.* (2010). "Conformational flexibility and allosteric regulation of cathepsin K." Biochem J **429**(2): 379-389.

Novinec, M., *et al.* (2014). "Probing the activity modification space of the cysteine peptidase cathepsin K with novel allosteric modifiers." PLoS One **9**(9): e106642.

Novinec, M., *et al.* (2014). "Cysteine cathepsin activity regulation by glycosaminoglycans." Biomed Res Int **2014**: 309718.

Novinec, M., *et al.* (2016). "An allosteric site enables fine-tuning of cathepsin K by diverse effectors." FEBS Lett **590**(24): 4507-4518.

Nussinov, R. (2016). "Introduction to Protein Ensembles and Allostery." Chem Rev **116**(11): 6263-6266.

Nussinov, R. and C. J. Tsai (2012). "The different ways through which specificity works in orthosteric and allosteric drugs." Curr Pharm Des **18**(9): 1311-1316.

Nussinov, R. and C. J. Tsai (2013). "Allostery in disease and in drug discovery." Cell **153**(2): 293-305.

Nussinov, R. and C. J. Tsai (2015). "Allostery without a conformational change? Revisiting the paradigm." Curr Opin Struct Biol **30**: 17-24.

Ou-Yang, S. S., *et al.* (2012). "Computational drug discovery." Acta Pharmacol Sin **33**(9): 1131-1140.

Panjkovich, A. and X. Daura (2012). "Exploiting protein flexibility to predict the location of allosteric sites." BMC Bioinformatics **13**: 273.

Papaleo, E., *et al.* (2016). "The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery." Chemical Reviews **116**(11): 6391-6423.

Pena, I., *et al.* (2015). "New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: an open resource." Sci Rep **5**: 8771.

Purdy, M. D., *et al.* (2014). "Function and dynamics of macromolecular complexes explored by integrative structural and computational biology." Curr Opin Struct Biol **27**: 138-148.

Rarey, M., *et al.* (1996). "A fast flexible docking method using an incremental construction algorithm." J Mol Biol **261**(3): 470-489.

Rastelli, G., *et al.* (2014). "Structure-based discovery of the first allosteric inhibitors of cyclin-dependent kinase 2." Cell Cycle **13**(14): 2296-2305.

Rathore, R. S., *et al.* (2013). "Advances in binding free energies calculations: QM/MM-based free energy perturbation method for drug design." Curr Pharm Des **19**(26): 4674-4686.

Rawal, R. K., *et al.* (2012). "Structure-activity relationship studies on clinically relevant HIV-1 NNRTIs." Curr Med Chem **19**(31): 5364-5380.

Ribeiro, A. A. S. T. and V. Ortiz (2016). "A Chemical Perspective on Allostery." Chemical Reviews **116**(11): 6488-6502.

Ribeiro, I., *et al.* (2009). "New, improved treatments for Chagas disease: from the R&D pipeline to the patients." PLoS Negl Trop Dis **3**(7): e484.

Rivalta, I., *et al.* (2012). "Allosteric pathways in imidazole glycerol phosphate synthase." Proc Natl Acad Sci U S A **109**(22): E1428-1436.

Roe, D. R. and T. E. Cheatham, 3rd (2013). "PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data." J Chem Theory Comput **9**(7): 3084-3095.

Rogers, K. E., *et al.* (2012). "Novel cruzain inhibitors for the treatment of Chagas' disease." Chem Biol Drug Des **80**(3): 398-405.

Ryckaert, J.-P., *et al.* (1977). "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes." Journal of Computational Physics **23**(3): 327-341.

Saalau-Bethell, S. M., *et al.* (2012). "Discovery of an allosteric mechanism for the regulation of HCV NS3 protein function." Nat Chem Biol **8**(11): 920-925.

Sajid, M. and J. H. McKerrow (2002). "Cysteine proteases of parasitic organisms." Mol Biochem Parasitol **120**(1): 1-21.

Sajid, M., *et al.* (2011). "Cruzain : the path from target validation to the clinic." Adv Exp Med Biol **712**: 100-115.

Salomon-Ferrer, R., *et al.* (2013). "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald." J Chem Theory Comput **9**(9): 3878-3888.

Schneider, G. (2010). "Virtual screening: an endless staircase?" Nat Rev Drug Discov **9**(4): 273-276.

Schneider, T. and E. Stoll (1978). "Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions." Physical Review B **17**(3): 1302.

Sethi, A., *et al.* (2009). "Dynamical networks in tRNA:protein complexes." Proc Natl Acad Sci U S A **106**(16): 6620-6625.

Shen, A. (2010). "Allosteric regulation of protease activity by small molecules." Mol Biosyst **6**(8): 1431-1443.

Skjaerven, L., *et al.* (2014). "Integrating protein structural dynamics and evolutionary analysis with Bio3D." BMC Bioinformatics **15**: 399.

Smooker, P. M., *et al.* (2010). "Cathepsin B proteases of flukes: the key to facilitating parasite control?" Trends Parasitol **26**(10): 506-514.

Stank, A., *et al.* (2016). "Protein Binding Pocket Dynamics." Acc Chem Res **49**(5): 809-815.

Stanley, N. and G. De Fabritiis (2015). "High throughput molecular dynamics for drug discovery." In Silico Pharmacol **3**: 3.

Steverding, D. (2014). "The history of Chagas disease." Parasit Vectors **7**: 317.

Swami, A. and B. Sadler (1997). Modulation classification via hierarchical agglomerative cluster analysis. Signal Processing Advances in Wireless Communications, First IEEE Signal Processing Workshop on, IEEE.

Tiberti, M., *et al.* (2014). "PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins." J Chem Inf Model **54**(5): 1537-1551.

Trossini, G. H., *et al.* (2009). "Quantitative structure-activity relationships for a series of inhibitors of cruzain from *Trypanosoma cruzi*: molecular modeling, CoMFA and CoMSIA studies." J Mol Graph Model **28**(1): 3-11.

Trott, O. and A. J. Olson (2010). "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading." J Compu Chem **31**(2): 455-461.

Tsai, C. J., *et al.* (2008). "Allostery: absence of a change in shape does not imply that allostery is not at play." J Mol Biol **378**(1): 1-11.

Tsai, C. J., *et al.* (1999). "Folding funnels, binding funnels, and protein function." Protein Sci **8**(6): 1181-1190.

Tsai, C. J. and R. Nussinov (2014). "A unified view of "how allostery works"." PLoS Comput Biol **10**(2): e1003394.

Turk, B. (2006). "Targeting proteases: successes, failures and future prospects." Nat Rev Drug Discov **5**(9): 785-799.

Turk, V., *et al.* (2012). "Cysteine cathepsins: from structure, function and regulation to new frontiers." Biochim Biophys Acta **1824**(1): 68-88.

van den Bedem, H. and J. S. Fraser (2015). "Integrative, dynamic structural biology at atomic resolution--it's about time." Nat Methods **12**(4): 307-318.

Van Wart, A. T., *et al.* (2014). "Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis." J Chem Theory Comput **10**(2): 511-517.

Verkhivker, G. M., *et al.* (2009). "Structural and computational biology of the molecular chaperone Hsp90: from understanding molecular mechanisms to computer-based inhibitor design." Curr Top Med Chem **9**(15): 1369-1385.

Verlet, L. (1967). "Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules." Physical review **159**(1): 98.

Verma, S., *et al.* (2016). "Cysteine Proteases: Modes of Activation and Future Prospects as Pharmacological Targets." Front Pharmacol **7**: 107.

Vettoretti, G., *et al.* (2016). "Molecular Dynamics Simulations Reveal the Mechanisms of Allosteric Activation of Hsp90 by Designed Ligands." Sci Rep **6**: 23830.

Wagner, J. R., *et al.* (2016). "Emerging Computational Methods for the Rational Discovery of Allosteric Drugs." Chemical Reviews **116**(11): 6370-6390.

Wang, J., *et al.* (2006). "Automatic atom type and bond type perception in molecular mechanical calculations." J Mol Graph Model **25**(2): 247-260.

Wang, J., *et al.* (2004). "Development and testing of a general amber force field." J Comput Chem **25**(9): 1157-1174.

Wenthur, C. J., *et al.* (2014). "Drugs for allosteric sites on receptors." Annu Rev Pharmacol Toxicol **54**: 165-184.

Wereszczynski, J. and J. A. McCammon (2012). "Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition." Q Rev Biophys **45**(1): 1-25.

Wild, C., *et al.* (2014). "Allosteric Modulation of G Protein-Coupled Receptors: An Emerging Approach of Drug Discovery." Austin J Pharmacol Ther **2**(1).

Wilkinson, S. R. and J. M. Kelly (2009). "Trypanocidal drugs: mechanisms, resistance and new targets." Expert Rev Mol Med **11**: e31.

Wu, P., *et al.* (2015). "FDA-approved small-molecule kinase inhibitors." Trends Pharmacol Sci **36**(7): 422-439.

Yuriev, E., *et al.* (2015). "Improvements, trends, and new ideas in molecular docking: 2012-2013 in review." J Mol Recognit.

Zeller, F. and M. Zacharias (2014). "Evaluation of Generalized Born Model Accuracy for Absolute Binding Free Energy Calculations." J Phys Chem B.

Zhao, H. and A. Caflisch (2015). "Molecular dynamics in drug design." Eur J Med Chem **91**: 4-14.

Zoete, V. and O. Michielin (2007). "Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein-protein association using MM-GBSA: application to the TCR-p-MHC complex." Proteins **67**(4): 1026-1047.

7. Appendix

Table A1. Vina score and effective free energy values of 60 hits obtained against cruzain site 3.

	Compound ID	S_{vina} (kcal/mol)	ΔG_{eff}^{a} (kcal/mol)
Cluster 1	ZINC58209662	-7.80	-23.82
	ZINC71863347	-7.80	-23.40
	ZINC23128396	-7.90	-22.76
	ZINC23956941	-8.00	-20.56
	ZINC71886094	-7.70	-19.20
	ZINC06371005	-7.70	-18.24
	ZINC06370772	-7.70	-17.70
	ZINC95476996	-7.80	-17.35
	ZINC49383246	-7.70	-17.24
	ZINC67283777	-7.70	-16.19
	ZINC12916000	-7.90	-16.15
	ZINC97083511	-7.80	-15.28
	ZINC04497409	-8.00	-14.21
	ZINC64986957	-7.70	-14.19
	ZINC08695467	-7.70	-14.14
	ZINC00233557	-7.80	-9.12
	ZINC72290405	-7.70	-8.87
	ZINC00233556	-7.70	-8.81
	ZINC08695469	-7.70	-6.69
	ZINC89851097	-7.80	-4.66
Cluster 3	ZINC40466547	-7.00	-19.38
	ZINC23192646	-7.00	-17.98
	ZINC40773138	-7.00	-17.75
	ZINC40790379	-7.10	-17.65
	ZINC47320570	-7.00	-17.33
	ZINC40773248	-7.00	-17.27
	ZINC20746082	-7.20	-17.07
	ZINC70039563	-7.10	-15.73
	ZINC00187790	-7.10	-15.33
	ZINC67842825	-7.20	-15.31
	ZINC25063292	-7.50	-15.00
	ZINC10024043	-7.00	-14.66
	ZINC48265141	-7.00	-14.40
	ZINC95360982	-7.70	-14.40
	ZINC00489137	-7.10	-14.27
	ZINC40789993	-7.00	-14.06
	ZINC46577698	-7.60	-12.95
	ZINC21891395	-7.00	-11.05

	ZINC21707700	-7.20	-10.95
	ZINC20140709	-7.10	-8.99
Cluster 4	ZINC00352089	-9.10	-41.05
	ZINC83627668	-9.10	-31.02
	ZINC17322062	-9.40	-25.21
	ZINC95426508	-9.40	-22.53
	ZINC21891395	-8.90	-20.91
	ZINC21219885	-9.10	-20.17
	ZINC14319810	-9.00	-19.80
	ZINC71863529	-9.00	-19.33
	ZINC82710332	-8.90	-18.97
	ZINC44895608	-8.90	-18.38
	ZINC19145494	-8.90	-18.11
	ZINC24958873	-8.90	-17.77
	ZINC32883757	-9.10	-16.76
	ZINC82710358	-8.90	-16.60
	ZINC12901838	-8.90	-16.46
	ZINC09344321	-8.90	-15.85
	ZINC05799926	-9.20	-15.82
	ZINC23609398	-9.00	-14.57
	ZINC29790160	-9.00	-13.76
	ZINC05303128	-9.00	-10.98

[a] Effective free energy $\Delta G_{eff} = \Delta E_{MM} + \Delta G_{sol}$

Table A2. Vina score and effective free energy values of 60 hits obtained against cruzain site 1.

	Compound ID	**S_{vina} (kcal/mol)**	**ΔG_{eff}^{a} (kcal/mol)**
Cluster 1	ZINC25063292	-7.10	-15.05
	ZINC69595267	-6.80	-11.37
	ZINC67455750	-6.70	-17.65
	ZINC69457407	-6.70	-16.82
	ZINC76093363	-6.70	-21.01
	ZINC84553407	-6.70	-23.42
	ZINC91032079	-6.70	-14.35
	ZINC40023366	-6.60	-21.09
	ZINC09120852	-6.60	-24.52
	ZINC71853953	-6.60	-15.00
	ZINC92224428	-6.60	-13.34
	ZINC00527137	-6.60	-13.25
	ZINC32064738	-6.60	-20.49
	ZINC05700314	-6.50	-16.36
	ZINC44910097	-6.50	-17.14
	ZINC44910108	-6.50	-19.53
	ZINC58901535	-6.50	-8.30
	ZINC67455749	-6.50	-6.62

	ZINC69776120	-6.50	-23.16
	ZINC84089991	-6.50	-14.23
Cluster 6	ZINC01394580	-7.00	-15.37
	ZINC69595267	-7.00	-10.31
	ZINC46577698	-6.80	-11.62
	ZINC13006207	-6.70	-13.76
	ZINC20859600	-6.70	-18.74
	ZINC65503506	-6.70	-14.29
	ZINC95477015	-6.70	-16.46
	ZINC03082372	-6.60	-14.21
	ZINC72412287	-6.60	-6.56
	ZINC01461504	-6.60	-10.88
	ZINC05069514	-6.60	-14.69
	ZINC65534408	-6.60	-17.52
	ZINC67455749	-6.60	-16.35
	ZINC75532874	-6.60	-15.54
	ZINC01227557	-6.50	-8.79
	ZINC06423848	-6.50	-11.29
	ZINC12527386	-6.50	-11.21
	ZINC22415199	-6.50	-13.94
	ZINC29374686	-6.50	-12.79
	ZINC40773330	-6.50	-14.83
Cluster 7	ZINC89499499	-6.90	-12.93
	ZINC01394580	-6.80	-20.28
	ZINC08393816	-6.80	-17.34
	ZINC00529177	-6.80	-14.11
	ZINC05313659	-6.80	-13.79
	ZINC05313664	-6.80	-14.34
	ZINC06647942	-6.80	-6.08
	ZINC65109544	-6.80	-14.96
	ZINC95426252	-6.80	-20.12
	ZINC00264656	-6.70	-15.21
	ZINC12407436	-6.70	-12.60
	ZINC22010820	-6.70	-14.24
	ZINC95453749	-6.70	-6.65
	ZINC04546305	-6.70	-13.29
	ZINC05531161	-6.70	-15.03
	ZINC20859600	-6.70	-17.25
	ZINC40790034	-6.70	-17.00
	ZINC40790227	-6.70	-8.09
	ZINC48080948	-6.70	-9.06
	ZINC67757204	-6.70	-16.27

[a] Effective free energy $\Delta G_{eff} = \Delta E_{MM} + \Delta G_{sol}$

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

TERMO DE REPRODUÇÃO XEROGRÁFICA

Autorizo a reprodução xerográfica do presente Trabalho de Conclusão, na íntegra ou em partes, para fins de pesquisa.

São José do Rio Preto, 29 de Setembro de 2017.

Lilian Hernández Alvarez