UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

SCUOLA DI DOTTORATO IN SCIENZE E TECNOLOGIE CHIMICHE

DIPARTIMENTO DI CHIMICA

DOTTORATO IN SCIENZE CHIMICHE XXVI CICLO

# COMPUTATIONAL MODELLING OF BIOMOLECULAR SYSTEMS: APPLICATIONS TO THE STUDY OF MOLECULAR RECOGNITION PROCESSES

Fabio Doro

N. R09033

Tutor: Dr. Laura Belvisi

Co-tutor: Dr. Monica Civera

Coordinator: Prof. Emanuela Licandro

A.A. 2012-2013

# Table of Contents

**INTRODUCTION AND METHODS**

# INTRODUCTION AND METHODS

# Chapter 1: Introduction

> Today the computer is just as important a tool for chemists as the test tube. Simulations are so realistic that they predict the outcome of traditional experiments.
>
> Nobel Prize in Chemistry 2013, Press Release[1]

On October 9, 2013 the Nobel Prize in Chemistry was awarded to Martin Karplus, Michael Levitt and Arieh Warshel for “the development of multiscale models for complex chemical systems”. As the Royal Swedish Academy of Sciences noted “Chemists used to create models of molecules using plastic balls and sticks. Today, the modelling is carried out in computers...Computer models mirroring real life have become crucial for most advances made in chemistry today.”

Molecular modelling has been defined as "the compendium of methods for mimicking the behavior of molecules or molecular systems". [2] It has been successfully used in several research fields. For example, if we look at the enormous advances of biochemistry over the last 50 years, we see that huge technological advances have taken place in sequencing single biomolecules, in mapping structure and dynamics via Electron Microscopy (EM), X-ray diffraction and Nuclear Magnetic Resonance (NMR), and more. It is worth noting that in order to analyze the spin-spin coupling obtained from a NMR experiment, or to analyze the diffraction pattern of an X-ray, modelling is needed. Often, the information given by the X-ray or NMR

[1] http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/press.html

[2] H. Senderowitz, http://chem.yu.edu.jo/rawash/chem_495/MM1.ppt

experiment is not conclusive and does not allow a unique determination of the structure of the sample. Molecular modelling is then used to calculate the energy of the structure based on theoretical and empirical potentials describing the energy of the system. Moreover, a usual step in the interpretation of the diffraction data of a biomolecule is the refinement step, in which molecular modelling techniques are used to better associate the electron density map to each individual atoms. In the reaction mechanism studies, theoretical methods can be applied to search for the transition state structure, that is not experimentally visible.

As the advancing experimental techniques give more in-depth insight into the static properties of proteins and nucleotides, the encompassed ability of molecular modelling of understanding their dynamics still stands.

Moreover, when experimental techniques show difficulties in the characterization of molecular movements, for example the time scales involved are too short for any experimental measure, or the conditions necessary to perform the experiment could not be obtained, computational modelling studies can provide reliable information on these structural variations.

Biomolecules are probably one of the most complex subject to study considering both the high number of degrees of freedom per system (composed by macromolecules and surrounding environment) and the huge hierarchy of functionally significant timescales movements, which can vary from nanoseconds to milliseconds and beyond.

Considering the problem from a molecular point of view, several methods based on classical physics have been developed in order to characterize macromolecules and their interactions within biological environments.

In this thesis a broad range of computational modelling techniques, which will be discussed in chapter 2, has been used to study biomolecular systems. In particular two main topics have been addressed:

- the conformational analysis of unnatural glycopeptides, in which Monte Carlo methods have been used to sample different molecule conformations in order to analyze and characterize their structural, dynamical and functional properties (chapters 3 and 4).
- the analysis of protein-protein interactions in classical cadherins and the design of small peptidomimetic inhibitors (remaining chapters). Here, molecular dynamics techniques (both biased and unbiased) together with computer-aided drug design were used to study the structural properties and the mechanism of the cadherin homophilic binding and to design the first class of small molecule inhibitors of their interaction.

The research activities described in the thesis have led to the following publications and communications:

**Publications**

- α-N-Linked glycopeptides: conformational analysis and bioactivity as lectin ligands.; Marcelo, F.; Cañada, F. J.; André, S.; Colombo, C.; Doro, F.; Gabius, H. J.; Bernardi, A.; Jiménez-Barbero, J.; *Org. Biomol. Chem.* **2012**, *10* (30), 5916-5923, DOI: 10.1039/C2OB07135E.
- Design of novel peptidomimetic inhibitors of cadherin homophilic interactions; Doro, F.; Colombo, C.; Alberti, C.; Arosio, D.; Belvisi, L.; Casagrande, C.; Fanelli, R.; Manzoni, L.; Piarulli, U.; Tomassetti, A.; Civera, M.; *submitted to Chem. Eur. J.*
- Reconstructing the free energy landscape of E-cadherin conformational transition by atomistic simulations.; Doro, F.; Saladino, G.; Belvisi, L.; Civera, M.; Gervasio, F. L.; *manuscript in preparation*

**Communications**

- Conformational Analysis and Molecular Dynamics simulations of α-N-linked glycopeptides, Doro, F.; Marcelo, F.; Colombo, C.; Stucchi, M.; Vasile, F.; Bernardi, A.; Jiménez-Barbero, J.; 26th International Carbohydrate Symposium, P155, Madrid July 22nd – 27th, **2012** (poster communication)
- Design and synthesis of peptidomimetic molecules targeting cadherin-mediated protein-protein interactions, Colombo, C.; Alberti, C.; Arosio, D.; Belvisi, L.; Doro, F.; Manzoni, L.; Civera, M.; Ischia Advanced School of Organic Chemistry (IASOC 2012), P12, Ischia (Naples) September 22nd – 26th, **2012** (poster communication)
- Computer-aided design of peptidomimetic molecules targeting cadherin-mediated protein-protein interactions, Doro, F.; Belvisi, L.; Civera, M.; 2nd National Meeting Computationally Driven Drug Discovery, Genova, February 4th-6th, **2013** (oral communication)
- Modelling cadherin-mediated protein-protein interactions by atomistic simulations, Doro, F.; Belvisi, L.; Saladino, G.; Gervasio, F. L.; Civera, M.; 5th European Conference on Chemistry for Life Sciences, Barcelona, June 10th – 12th, **2013** (Abstract Book P-083, poster communication)

# Chapter 2: Methods

The theoretical framework of this thesis, based on classical molecular mechanics, is described in this chapter. Methods used in the characterization of biomolecules dynamics and equilibrium (Molecular Dynamics and Monte Carlo simulation, conformational analysis and metadynamics) and in the drug discovery process (computational alanine scanning, 3D database searching and molecular docking) are discussed. A more detailed discussion can be found in specific textbooks.[1,2]

## 2.1 MOLECULAR MECHANICS

In a quantum mechanical description of a molecule, electrons need to be included. Thus a very large number of particles must be considered in order to fully describe the system. Calculations performed at a quantum mechanical level are very time consuming and a different approach is needed when dealing with molecules bigger than a few tens of atoms. Molecular mechanics is then invariably used in this case. Molecular mechanics is based on the validity of two main assumptions:

- the Born-Oppenheimer approximation and the Potential Energy Surface
- Force fields

While the Born-Oppenheimer approximation enables the possibility of writing the potential energy of the system as a function of the nuclear coordinates, discarding the electrons, the force fields give a functional form to the description of the potential energy.

### 2.1.1 The Born-Oppenheimer approximation and the Potential Energy Surface

In describing the molecular system one wants to separate the motion of atoms into time-independent electron and time-dependent atomic nuclei motion.

The time-dependent Schrödinger equation describes the time evolution of a quantum system.

$$H\psi = i\hbar\frac{\delta\psi}{\delta t} \qquad (1)$$

The Hamiltonian H describes the sum of potential and kinetic energy. ψ is the wave function which contains the information about all the particles of the system, ħ is the Planck constant over 2π and *i* is the imaginary unit.

Since the rigorous calculation of the solution of the Schrödinger equation for multiple nuclei and electrons is not possible, Max Born and J. Robert Oppenheimer intuition was to decouple the motion of the atomic nuclei from the motion of the electrons.[3] If the speed of the atomic nuclei is small compared to that of the electrons, it can be assumed with good approximation that the electrons adapt instantaneously to the nuclear configuration. So, the wave function Ψ of the complete system can be expressed as the product of the time-dependent wave function of the atomic nuclei $\Psi_n$ and the time-independent electron wave function $\Psi_e$.

$$\Psi(\mathbf{R}, \mathbf{r}, t) = \Psi_n(\mathbf{R}, t)\Psi_e(\mathbf{r}; \mathbf{R}) \qquad (2)$$

where $\boldsymbol{R} = (\boldsymbol{R}_1, \boldsymbol{R}_2, .., \boldsymbol{R}_n)$ and $\boldsymbol{r} = (\boldsymbol{r}_1, \boldsymbol{r}_2, .., \boldsymbol{r}_K)$ are the coordinates of the N nuclei and the K electrons, respectively.

The wave function $\Psi_e$ of the electronic state is only dependent on the coordinates **R** and not on the velocities $\mathbf{v_R}$ of the nuclei. Fixed nuclear positions **R** can then be used to solve the equation

$$H_e(\mathbf{R})\Psi_e(\mathbf{r}; \mathbf{R}) = E_e(\mathbf{R})\Psi_e(\mathbf{r}; \mathbf{R}) \qquad (3)$$

The Hamiltonian $H_e = H - H_n$ is simply the complete Hamiltonian removed of the Hamiltonian of the nuclei $H_n$. $E_e(\boldsymbol{R})$ of eq. (3) are the energy eigenvalues. By applying (2) and (3) to (1), the final time-dependent Schrödinger equation for the nuclei is obtained:

$$(T_n + E_e(\mathbf{R}))\Psi_n(\mathbf{R},t) = i\hbar \frac{\delta\Psi_n(\mathbf{R},t)}{\delta t} \tag{4}$$

$T_n$ is the kinetic energy of the nuclei. The Born-Oppenheimer approximation will not break down as long as the eigenvalues $E_e$ of eq. (3) do not overlap. This is usually the case for molecules in the ground state.

However, even in the case where the Born-Oppenheimer approximation is valid, solving the electronic Schrödinger equation (3) or the time-dependent Schrödinger equation for the nuclei (4) is still not feasible for big molecules. Thus nuclei are considered as point particles following classical, Newtonian mechanics. Energy changes are then associated to movements of the nuclei on a Potential Energy Surface (PES), which corresponds to the electronic ground state eigenvalue $E_e^0(\mathbf{R})$:

$$V(\mathbf{R}) = E_e^0(\mathbf{R}) \tag{5}$$

### 2.1.2 Force fields

The problem now consists of finding a suitable expression for $V(\mathbf{R})$. A typical potential function takes the form of a sum of classical potential energy expressions, each tunable using adjustable parameters.

$$V(\mathbf{R}) = V_{bonds} + V_{angles} + V_{dihedrals} + V_{pairs} \tag{6}$$

Each term can be expressed using specific functions, such as:

$$\begin{aligned} V(\mathbf{R}) = &\sum_{bonds} \frac{k_b}{2}(l - l_{eq})^2 \\ &+ \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_{eq})^2 \\ &+ \sum_{dihedrals} \frac{V_n}{2}(1 + \cos(n\Phi - \delta)) \\ + \sum_{pairs\, i,j} 4\epsilon_{ij}&\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 R_{ij}} \end{aligned} \tag{7}$$

Figure 1 illustrates the main energy terms contributing to a force field $V(\mathbf{R})$ having the functional form of eq. (7). Each term describes a pairwise relation, taking into account the physico-chemical interactions present in the system. The bond and angle force field parameters are the force constants $k_b$, $k_\theta$, the equilibrium angles $\theta_{eq}$ and equilibrium bond lengths $l_{eq}$. The torsion potential is described by its multiplicity $n$, its barrier height $V_n$ and its phase $\delta$. Improper dihedrals are treated analog to normal angles. The non-bonded parameters are the partial charges $q_i$ and the Lennard-Jones (LJ) parameters $\sigma_{ij}$ and $\varepsilon_{ij}$.

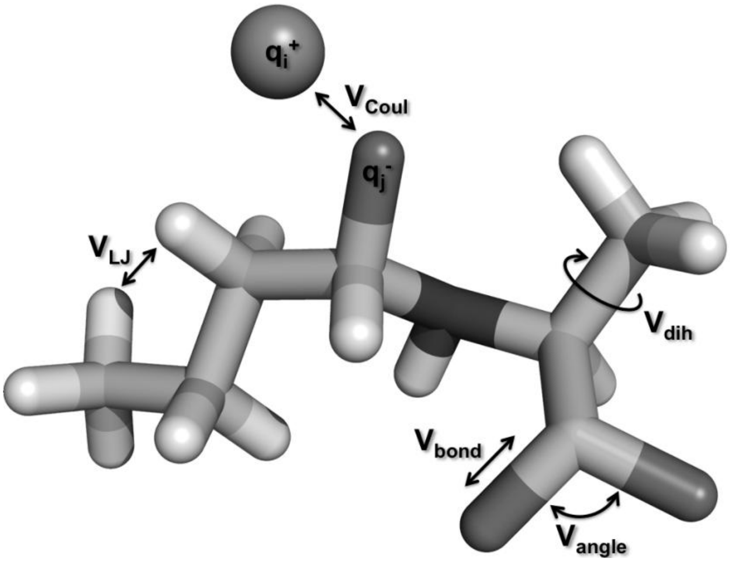


Figure 1. A common set of force field terms describing the potential energy surface of a molecule. Both bonded ($V_{bond}$, $V_{dih}$, $V_{angle}$) and non bonded potentials ($V_{LJ}$, $V_{Coul}$) are shown.

Several parameters are present in the force field, such as the equilibrium distances, force constants or Van der Waals and electrostatic terms, all of them have to be determined by either using experimental data or through the fitting to high level ab initio calculations. Clearly, the parametric nature of the force fields imposes restrictions to their uses in contexts which are different from the ones they have been developed for, so for example it would be unwise to use a force field set for amino acids in a inorganic polymer study. Hence, a list of various force fields has been developed, and it is constantly being upgraded. The existing force fields include, among others, AMBER,[4,5] CHARMM,[6,7] GROMOS,[8] MM2-4,[9–11] MMFF[12] and OPLS.[13]

## 2.2 MOLECULAR DYNAMICS

Molecular Dynamics (MD) simulations are based on the calculation of the time dependent behavior of molecular systems, giving a detailed description of the variation from one conformation to another of the system studied. Simulations generate ensembles of representative configurations in such a way that accurate values of thermodynamic and structural properties can be obtained with a reasonable amount of computation. In particular, statistical analysis links microscopic and macroscopic properties providing the fundamental principles for the description of biomolecular systems.

### 2.2.1 Time and Ensemble Averages

In general, macroscopic properties such as pressure, heat capacity, volume etc depend on the positions and momenta of the N particles constituting the system. The value of a particular property A at a certain time $t$ can be defined as a function of $\mathbf{p}^N(t)$ and $\mathbf{R}^N(t)$ representing the $N$ momenta and positions of the particles at time t, respectively. The instantaneous value of A can thus be written as:

$$A(t) = A\left(\mathbf{p}^N(t), \mathbf{R}^N(t)\right) \qquad (8)$$

that is as a function of all the positions and momenta of the particle at time $t$. During time, the values of quantity $A$ change under the effect of temperature fluctuations and interactions between particles. Experimentally, it is impossible to measure the single value of $A$ at time $t$, but it is possible to measure the average of $A$ during the time in which the measurement is carried out, and therefore it represents a time average. As the time over which the measurement is made grows, the average value of A approaches its real equilibrium value. The average value of $A_{ave}$ can thus be written as:

$$A_{ave} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A\big(\mathbf{p}^N(t), \mathbf{R}^N(t)\big)\, dt \tag{9}$$

In this case, the time of the measurement is much longer than the typical relaxation time of each event and the average value represents the equilibrium one.

If one has an energy function (i.e. a force field) which describes the interactions between the particles and a way of calculating the forces acting on the particles (i.e using Newton laws), then the positions and momenta of all particles could be computed for every *t*. Applying equation (9) would then provide the average values of the property of interest. Unfortunately, the dimensions of real molecular systems are so that it is impossible to calculate all the interactions for a number of particles of the order of $10^{23}$.

This problem can be overcame with statistical mechanics. In statistical mechanics the attention is not focused just on one single system evolving in time, but rather on a large number of replicas of the same system evolving simultaneously. As a consequence the time average is replaced by an ensemble average:

$$\langle A \rangle = \iint d\mathbf{p}^N d\boldsymbol{R}^N A(\mathbf{p}^N, \boldsymbol{R}^N)\rho(\mathbf{p}^N, \boldsymbol{R}^N), \tag{10}$$

Here, $\langle A \rangle$ corresponds to the ensemble average or expectation value of property *A*, i.e. the average value of *A* over all the replicas of the system in the ensemble generated by the simulation. $\rho(\mathbf{p}^N, \boldsymbol{R}^N)$ is the probability density of the ensemble, meaning the probability to obtain a configuration with momenta $\mathbf{p}^N$ and positions $\mathbf{R}^N$ among all the configurations sampled in the simulation. If the simulation is long enough to sample all the relevant configurations for the system for the ergodic hypothesis the ensemble average will be equivalent to the time average. Under these conditions the density of probability is described by the typical Boltzmann distribution:

$$\rho(\mathbf{p}^N, \boldsymbol{R}^N) = \frac{1}{Q}\mathrm{e}^{-\mathrm{E}(\mathbf{p}^N, \boldsymbol{R}^N)/\mathrm{k_B T}} \tag{11}$$

where E is the energy function, *Q* the partition function, $\mathrm{k_B}$ the Boltzmann's constant and *T* the temperature. The partition function is generally written in terms of the Hamiltonian *H* governing the system, e.g.

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}}\mathrm{e}^{-\mathrm{H}(\mathbf{p}^N, \boldsymbol{R}^N)/\mathrm{k_B T}} \tag{12}$$

The subscript *NVT* indicates a systems with a constant volume *V*, number of particles *N* and temperature *T* (Canonical Ensemble). In MD simulations on biological systems the Hamiltonian *H* can be approximately considered equal to the total energy *E* of the system. *N!* arises from the fact that particles are not distinguishable while $1/h^{3N}$ is related to the equivalence of the partition function to that calculated through quantum mechanics.

MD simulations generate a trajectory consisting of a collection of subsequent configurations and describing how the dynamic variables vary in the time. Thermodynamic quantities are calculated from the trajectory using numerical integration of equation (10).

### 2.2.2 Trajectory calculation

The configurations composing the trajectory of the system are generated through the application of Newton's laws of motion:

$$\mathbf{F}_i = m_i\mathbf{a}_\mathrm{i} \tag{13}$$

$$-\nabla_{\mathrm{R_i}}V(\mathbf{R}) = \frac{\delta^2\mathbf{R}_i}{\delta t^2} \tag{14}$$

applied on every $i$th atom of the system. The progression of the system is computed in small time steps $\Delta t$, using, for example, the leap-frog algorithm:[14]

$$\mathbf{v}_i\left(\mathrm{t}+\frac{\Delta t}{2}\right)=\mathbf{v}_i\left(\mathrm{t}-\frac{\Delta t}{2}\right)+\frac{\mathbf{F_i}(t)}{m_i}\Delta\mathrm{t} \tag{15}$$

$$\mathbf{R_i}(t+\Delta t)=\mathbf{R_i}(\mathrm{t})+\mathbf{v_i}\left(\mathrm{t}+\frac{\Delta t}{2}\right)\Delta t \tag{16}$$

Solving these equations will give the positions $\mathbf{R}(t)$ and velocities $\mathbf{v}(t)$ for all atoms $i=1,2,\ldots,N$ in the system. The time step $\Delta\mathrm{t}$ of the simulation is usually restricted to 1 fs to take into account the bond and angle vibrations involving hydrogen atoms. If these vibrations are constrained, the time step can be increased up to 2-4 fs.

### 2.2.3 Simulation details

In this section, a more in depth information about performing MD simulations is given.

First, to run an MD simulation it is necessary to define a molecular system to study and to identify the properties useful for its characterization. Generally, the model system will consist of *N* particles which will interact under the action of the potential and forces defined in equation (7).

An MD trajectory is divided into two parts, the first one is the equilibration stage, in which the system (and the properties of interest) will evolve as a function of time, and the second one is the production phase where it is possible to carry out the effective measurements as the system has reached the equilibrium. The choices of the model, the equilibration time and the way the measurement is carried out are very sensitive points which have to be evaluated carefully. Indeed incorrect results or bad artifacts can be generated by using the wrong model to describe the phenomena, by using too short equilibration/measurement times, or by failing to notice irreversible and chemically meaningless changes that could occur in the system.

### *2.2.3.1 Starting conformation*

To start the simulations initial positions and velocities to all atoms in the system must be assigned. In the case of biological molecules or protein simulations, the initial positions can be obtained from structural determination experiments such as NMR or X-Ray measurements.

Clearly, in biomolecular simulations, the user is mostly interested in investigating the properties of the system in presence of the appropriate solvent (or mixture of solvents), rather than simply studying gas-phase properties. To this end, the solute (protein, DNA, drugs, etc. . . ) is inserted in a pre-equilibrated solvent bath (any solvent molecules whose coordinates are too close to the solute atoms are eliminated from the system). In theory a simulation should be able to reproduce the behavior of an infinite system or of a real system of around $10^{23}$ particles, in which a negligible number of particles would be in contact with the boundaries (like the vessel walls in real-life experiments), in order to calculate straightforwardly macroscopic quantities. In practice this situation is completely out of reach even for the most powerful computers, and the study has to be carried out on finite-size systems characterized by some boundaries.

The correct choice of the method to treat the boundaries of the simulation is fundamental for the calculation of the properties of interest.

### *2.2.3.2 Periodic boundary conditions*

Periodic Boundary Conditions (PBC) are useful to run a simulation considering a relatively small number of particles, in such a way that the particles experience interactions and forces as if they were in a bulk fluid. The simplest representation of such a system is represented by a cubic box of particles which is replicated in all directions to give a periodic array (Figure 2).

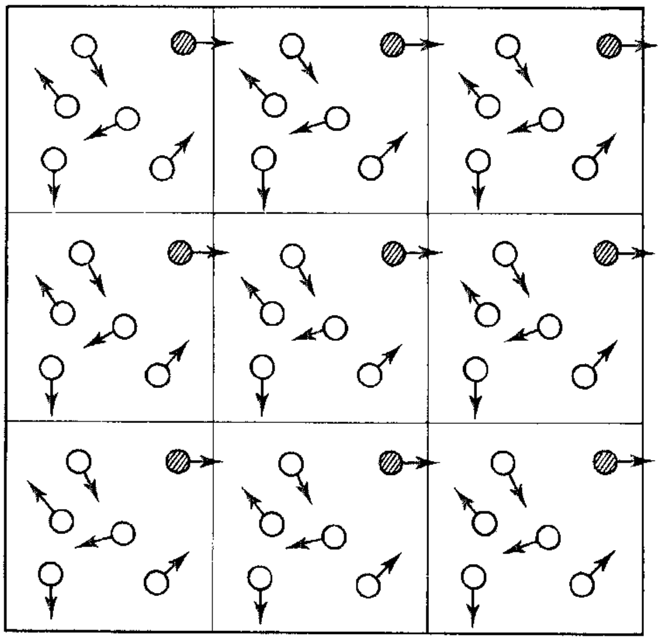


Figure 2. Periodic boundary conditions in two dimensions.

The particles coordinates in the replica images are obtained by adding to the original ones multiples of the box sides. If a particle leaves the box during the simulation, it will be replaced by its image coming in from the opposite side of the box. In this way the number of particles in the simulation box is kept constant and the solvent behaves basically as a bulk with no artifacts affecting the results of the simulation. To reduce the cost in term of calculation, periodic boundary conditions are most often combined with the minimum image convention: only one, the nearest-image of each particle is considered for short-range non-bonded interaction terms.

After the solvation of the system, initial atomic positions must be carefully checked to avoid any sizable overlap of groups. On average initial structures have to be minimized in order to remove bad contacts and optimize bond, angular and torsional interactions. The minimization procedure has the main objective to place the initial structure on low energy points on the Potential Energy Surface. For the starting movement, to each particle is assigned an initial velocity chosen from a uniform Maxwellian distribution of velocities consistent with the temperature at which the simulation will be run.

$$p(\boldsymbol{v}_i) = \sqrt{\frac{m_i}{2\pi kT}} e^{\frac{m_i v_i^2}{2kT}} \tag{17}$$

### *2.2.3.3 Solvation models*

In the most detailed microscopic approach, solvent molecules are treated explicitly, and the electrostatic properties of both solvent and solute are obtained by averaging over a very large number of configurations of the system. In the most widely used explicit model, TIP3P,[15] the water molecule is considered as a rigid molecule having three interaction sites, corresponding to its three atoms. The partial positive charges on the hydrogen atoms are exactly balanced by an appropriate negative charge located on the oxygen atom. The van der Waals interaction between two water molecules is computed using a Lennard-Jones function with just a single interaction point per molecule centered on the oxygen atom; no van der Waals interactions involving the hydrogen atoms are calculated.

The use of rigid molecules is of course an approximation. Even so, explicit solvation calculations of this kind run slowly because hundreds of explicit solvent molecules need to be included. This prompted interest in models which incorporate the influence of the solvent in an implicit fashion.[16] The simplest way of including solvent-related effects in molecular mechanics calculation is to increase the dielectric constant in the coulombic electrostatic term of the potential energy. Polar solvents, like water, dampen the electrostatic interactions and their effect can be modeled by assigning the appropriate dielectric constant value. Other methods, based on continuum electrostatics, define the solute interior and the solvent as regions with different dielectric constants, and the electrostatic solvation free energy is computed by solving the Poisson-Boltzmann equations.[17] These methods represent a rigorous treatment of continuum electrostatics, which takes into account solvation of single charges as well as screening effects (charge-charge and charge-dipole interactions). However, the calculations are too time-consuming to be performed routinely. To address this problem, empirical models have been developed. Here the solvation free energy is divided in three components, an electrostatic component, a van der Waals component and one component associated with creating the solute cavity within the solvent

$$\Delta G_{sol} = \Delta G_{elec} + \Delta G_{VdW} + \Delta G_{cav} \tag{18}$$

The last two terms depend on the solvent-accessible surface area of the solute and the first term is usually derived by the Generalized Born model.[18,19]

Implicit solvent models have advantages and disadvantages over explicit solvation models. Implicit methods allow inclusion of solvent effects at a fraction of the cost required by explicit models and do not generally show convergence problems. On the other hand, explicit solvation is much more efficient at handling charged groups and molecules. Furthermore, effects that may be related to the presence of individual water molecules in the vicinity of the solute such as the formation of solute-solvent hydrogen bonds, or in areas of molecular complexes that are not seen by the software as "solvent accessible" cannot be properly modeled by implicit solvation.

### *2.2.3.4 Non-bonded interactions*

The most time consuming part in an MD simulation is the calculation of non-bonded interactions, the $V_{pairs}$ term in eq. (6). As can be seen from eq. (7), the number of interactions scales with $O(N^2)$, as it is necessary to take into account the force contribution acting on particle i due to the presence of all its neighbors. To improve the scaling, fast decaying Lennard-Jones potentials can be cut off at around 10-15 Å.

However, the long ranging Coulomb interactions decay slowly with $R^{-1}$ and cutting them off would not be advisable.[20] Truncating the potentials in non natural ways leads to errors, especially in the cases of charged systems, where electrostatics is particularly important.

Therefore, other methods for treating long-range electrostatics, such as the Particle Mesh Ewald (PME) algorithm,[21,22] have been developed. In this algorithm the electrostatic potential is separated in two parts, one short range term, which is calculated in real space, and a long range term calculated in reciprocal space. By computing the long range part using a Fast Fourier Transform, a scaling of the algorithm with $O(NlogN)$ is achieved.

### *2.2.3.5 Temperature and pressure coupling*

To obtain an easier connection with experiments it is often desirable to run simulations at constant Temperature (T), and at constant Volume (V) or Pressure (P). The two most common simulation ensembles are in fact the NVT and NPT in which a fixed number of molecules, a constant temperature and, respectively, constant volume and pressure, are used.

Temperature is related to the average kinetic energy:

$$\langle K \rangle = \sum_{\text{atoms i}} \frac{m_i}{2} \boldsymbol{v}_i^2 = \frac{3}{2} N k_B T \qquad (19)$$

By changing the atom velocities at each step, it is possible to control the average kinetic energy and hence the temperature. More practically, one usually maintain the temperature close to the desired value by coupling the system to an external heat bath kept at the desired temperature. In the Berendsen algorithm[23] the bath acts as a heat reservoir which can supply or remove energy from the system. The velocities are scaled at each time step, such that the rate at which the temperature changes is proportional to the difference in temperature between the bath and the system.

The change in temperature is described by the following scaling formula:

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}(T_o - T(t)) \qquad (20)$$

Thus, the temperature deviation decays exponentially with a time constant $\tau$, and by changing $\tau$, the strength of the coupling can be varied and adapted to different situations. The main problem with this algorithm is that it does not generate rigorous canonical averages, since velocities are rescaled artificially. Depending on the scaling factor $\tau$, fluctuations between canonical and micro canonical ensembles are obtained.[24]

A new thermostat derived from the Berendsen algorithm, called velocity rescale,[25] has been shown to better sample the canonical distribution. Here a stochastic term is added to the equation describing the change in temperature:

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}\big(T_o - T(t)\big) + 2\sqrt{\frac{T_0 T(t)}{3N\tau_T}}\frac{dW}{d\mathrm{t}} \qquad (21)$$

where W is the function used for the Brownian motion.

Pressure fluctuations are generally much more pronounced since the pressure is related to the virial, which is obtained as the product of the positions and the derivatives of the potential energy function. This product is much more sensitive to the variations in position than the internal energy, which brings bigger pressure fluctuations. In any case, similarly with the temperature control, the system is coupled to an external pressure bath, with a scaling formula defined as:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}(P_o - P(t)) \qquad (22)$$

Again, $\tau$ is the time constant for the pressure coupling.

### *2.2.3.6 Structural properties*

Computer simulations allow the calculation of quantities which can be compared directly with experimental results (a fundamental step for the validation of simulation results) and also the prediction of properties inaccessible to experiments. Many different types of properties can be calculated ranging from average energies to structural and conformational information. Some structural properties relevant for the study of conformational variations and molecular interactions will be explained.

*Root Mean Square Deviation*

Root Mean Square Deviation (RMSD) generally contains information on the divergence in time of a structure from a reference one and it is determined with the following formula:

$$RMSD(t_1, t_2) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i \|r_i(t_1) - r_i(t_2)\|^2} \qquad (23)$$

Where M is the sum over all the atom masses, and $r_i$(t) is the position of atom *i* at time *t*. A protein is usually fitted on the backbone atoms N, Cα, C or just Cα atoms, but of course it is possible to compute the RMSD over other elements like only side chain atoms or even the whole protein. In general as reference structure for the calculation it is used the first one in the simulation (or a crystal one). In addition it can be defined a matrix with the RMSD as a function of $t_1$ and $t_2$, allowing easy identification of structural transitions in a trajectory.

*Radius of gyration*

The Radius of Gyration ($R_g$) gives a measure for the compactness of the structure and it is defined as:

$$R_g = \sqrt{\frac{\sum_i \|r_i\|^2 m_i}{\sum_i m_i}} \qquad (24)$$

This measure is very useful in the polymer field in order to describe the dimensions of a polymer chain; in MD simulations it gives information about the changes in the protein structure shape.

## 2.3 MONTE CARLO SIMULATION

In a Monte Carlo simulation, configurations of the system are generated by performing random changes to the atoms of the species of interest. For each configuration, the potential energy can be calculated, using only the positions of the atoms.

In principle, using a pure random search, the partition function of a system of *N* atoms, and thus the required thermodynamic properties, could be calculated using this simple algorithm:

1. Select a configuration of the system by randomly generating 3*N* Cartesian coordinates.
2. Calculate the potential energy of the configuration, $V(\mathbf{r}^N)$
3. From the potential energy, calculate the Boltzmann factor, $e^{-\frac{V(\mathbf{r}^N)}{k_B T}}$
4. Add the Boltzmann factor to the accumulated sum of Boltzmann factors and the potential energy contribution to its accumulated sum. Return to step 1
5. After a number $N_{trial}$ of iterations, the mean value of the potential energy would be calculated using:

$$\langle V(\mathbf{r}^N)\rangle = \frac{\sum_{i=1}^{N_{trial}} V_i(\mathbf{r}^N) e^{-V_i(\mathbf{r}^N)/k_B T}}{\sum_{i=1}^{N_{trial}} e^{-V_i(\mathbf{r}^N)/k_B T}} \qquad (25)$$

However, this approach is not feasible, since a large number of configuration have nearly zero Boltzmann factors. This reflects the nature of the phase space, most of which corresponds to non-physical configurations with very high energies. To get around this problem Metropolis et al[26] have found a way to generate configurations that make a large contribution to the integral. The crucial feature of the Metropolis approach is that it biases the generation of configurations towards those that make the most significant contribution to the integral. Specifically, it generates states with a probability $e^{-V_i(\mathbf{r}^N)/k_BT}$ and then counts each of them equally.

For many thermodynamic properties of a molecular system (one notable exception being the free energy), those states with a high probability p are also the ones that make a significant contribution to the integral.

Using a Monte Carlo Metropolis technique in conjunction with molecular dynamics has been attempted first by Guarnieri and Still.[27] Here, each stochastic dynamics step (SD, or velocity Langevin dynamics, as it adds a friction and a noise term to Newton's equations of motion) is followed by a Monte Carlo (MC) step (MC/SD). MC/SD performs constant temperature calculations that take advantage of the strengths of Metropolis Monte Carlo methods for quickly introducing large changes in a few degrees of freedom, and stochastic dynamics for its effective local sampling of collective motions. MC/SD does stochastic dynamics on the Cartesian space of a molecule and Monte Carlo on the torsion space of the molecule simultaneously. After each SD step a random deformation of some rotatable torsions is performed and accepted or rejected according to the Metropolis criteria. The next SD step is performed from the most recent configuration with the velocities taken from the previous SD step. The smooth merging of Monte Carlo and dynamics requires the use of the stochastic velocity Verlet integration scheme.

## 2.4 CONFORMATIONAL ANALYSIS

Conformational search methods are used to sample the PES, i.e. to locate all the accessible minima and their associated relative energies. Metropolis Monte Carlo and molecular dynamics simulation methods can be used to explore the entire conformational space of molecules. Methods which use a search algorithm coupled with an energy minimization are used to identify the preferred conformations of a molecule, i.e. the conformations localized at minimum points on the energy surface.

In the Monte Carlo / Energy Minimization (MC/EM) method, at each step the system randomly explores different regions of the PES by performing a random change of coordinates. Although the change can be made both in Cartesian coordinates or in internal coordinates, it has been shown that the latter method is much more efficient at exploring the conformational space of molecules, because it greatly reduces the number of degrees of freedom to be considered.[28] After the random change has been made, the structure is refined using energy minimization. If the minimized conformation has not been found before, and falls within a predefined energy window above the global minimum of the search, it is stored. A conformation is then selected to be used as the starting point for the next iteration, and the cycle starts again. The procedure continues until a given number of iterations have been performed or until the user decides that no new conformations can be found. There are many ways in which the structure for input to the next iteration can be selected, and this choice will influence the efficiency of the method. It has been shown that low-energy final conformers are favored by selecting low-energy starting geometries at each Monte Carlo step.[29]

### 2.4.1 Simulated annealing

At high temperatures, a system explores configurations of the phase space which are less probable and is able to easily overcome energy barriers. In simulated annealing, the temperature of the system is brought to high values and then gradually reduced.[30] Theoretically, at every T the system explores all the permitted conformations (using MC or MD techniques) and can reach the thermal equilibrium. With temperature lowering, states possessing less energy become favorable, according to the Boltzmann distribution. At T=0 K, the system should occupy only the state having the lowest energy, i.e. the global minimum of the PES. In practice, finding the global minimum would require an infinitesimal temperature gradient ($dT \rightarrow 0$) and at each temperature the system would need to reach the local minimum. Thus, several simulated annealing simulations are usually performed, obtaining a series of low-energy conformations.

## 2.5 FREE ENERGY CALCULATIONS: METADYNAMICS

Molecular dynamics and Monte Carlo simulations differ in many aspects. Molecular dynamics provides information about the time dependence of the properties of the system whereas successive Monte Carlo configurations are not time dependent. Molecular dynamics has a kinetic energy contribution to the total energy whereas in a Monte Carlo simulation the total energy is determined directly from the potential energy function. Nonetheless, both methods permit the calculation of a wide variety of thermodynamic properties, from the internal energy, to the heat capacity, to the radial distribution function. However, there are some properties, called ergodic or thermal properties, such as the free energy of the system, which cannot be easily derived. The reason is due to the fact that the configurations with a high energy make a significant contribution to the partition function $Q$ in the free energy expression $A = -k_B T lnQ$. The results for the free energy calculated using Monte Carlo or unbiased Molecular Dynamics methods, for a system bigger than a small molecule with fixed located minima, are then poorly converged and inaccurate. In the next section a method to estimate the free energy of a system, used in this thesis, is described.

### 2.5.1 Metadynamics

Metadynamics belongs to a family of methods called enhanced sampling techniques, in which the reconstruction of the probability distribution is in some way accelerated. The reconstructed probability distribution is function of one or a few predefined collective variables (CVs). A partial list of similar methods include thermodynamic integration,[31] free energy perturbation,[32] umbrella sampling,[33] weighted histogram techniques,[34] adaptive force bias,[35] and steered MD.[36] In metadynamics, in contrast to the other mentioned techniques, the system is disfavored to sample already visited regions through the addition of a repulsive Gaussian potential to the energy potential of the system. By doing so, a time-evolving bias contributes to the total Hamiltonian, making possible to cross energy barriers and finally obtaining a flat free energy surface.[37] During the last ten years, new metadynamics variants have contributed to evolve the method. One of these, the well-tempered metadynamics (WTM), is widely used because of the guaranteed convergence of the simulation. The method allows not only a faster diffusion but also the complete reconstruction of the underlying Free Energy Surface (FES) as a function of the chosen CVs.

During a metadynamics simulation, the total potential to which the system is subjected is due to the original potential $V(x)$ (the force field), plus a biasing term $V_B(s,t)$, $s$ being a small subset of CVs, all functions of the atomic coordinates of the systems. At regular time intervals $\tau$, repulsive Gaussians terms are added to the potential:

$$V_B(s,t) = \sum_{t=\tau,2\tau,\ldots} he^{-(s-s_t)^2/2\sigma^2} \tag{26}$$

where $\sigma^2$ and $h$ are the variance and the height of the gaussian, respectively. In a simulation described by this new potential, already visited states will have an energy penalty and will thus be less sampled. After all the free energy wells have been filled, the biasing potential will approximately corresponds to the original profile $F(s)$: [38]

$$V_B(s) \cong -F(s) \tag{27}$$

In well-tempered metadynamics the height of the Gaussian also is time dependent:

$$h(s,t) = h_0 \mathrm{e}^{-\frac{V_B(s,t)}{k_B \Delta T}} \tag{28}$$

As a consequence, the rate of the energy deposition, and thus the error, will tend to zero as the simulation proceeds. In WTM, $\Delta T$ of eq. (28) is added to the simulation temperature $T$ to give the fictitious $T + \Delta T$ higher temperature at which CVs are sampled.

The ratio $(T + \Delta T)/T$, called the bias factor, is typically used to define this fictitious temperature.

In order to obtain meaningful results, an appropriate set of CVs should be chosen. However, for a system with many degrees of freedom, one usually deals with non-optimal CVs. In this case, several methods can be used to speed the sampling along slow degrees of freedom. Indeed, the combination of the parallel tempering (PT) algorithm[39] with metadynamics (PT-MetaD) can be used for this purpose. In the replica exchange algorithm sampling is enhanced by exchange configurations of replicas simulated at different temperatures. More recently, in a new technique called Well-Tempered Ensemble[40] (WTE) the potential energy is used as the only CV. In the WTE one makes use of well-tempered metadynamics to obtain an estimate of the energy probability distributions at the various temperatures. This knowledge can then be used to implement larger energy fluctuations in subsequent PT or PT-MetaD simulations. By doing this, the number of replicas needed for the PT simulations are considerably reduced.

## 2.6 MOLECULAR MODELLING IN DRUG DISCOVERY

Molecular modelling techniques, such as molecular dynamics and conformational analysis, which have already been introduced, are widely used in drug discovery. Here I will briefly summarize a few other tools used in this thesis that do not fall in the previous sections. For a full review of these methods, see Ref. 1.

Usually in drug discovery, a small number of hit molecules (*hits*) is identified, to which a lead series (*leads*) follows. A hit is a molecule that has some reproducible activity in a biological test. A hit can be identified either by high-throughput screening (HTS), if the structure of the receptor is unknown, or by Structure Based Drug Design, if the structure of the receptor has already been characterized. Leads are a set of molecules structurally similar to the hit, showing differences in activity related to differences in structures. This leads to what is called lead optimization, the structural modification of the compound in order to enhance his activity. Molecular modelling can be used in all of these first steps of drug discovery, from the identification of the binding site of the receptor to the lead optimization of the hit molecule.

### 2.6.1 Computational alanine scanning

The mutation of specific residues in proteins has been long used to test the contribution of individual amino acids to the properties of proteins. Starting from the works of Hodges[41] and Fersht,[42] protein libraries have been collected to explore the relationship between the primary amino acid sequence and protein shape, stability and activity.

Alanine-scanning mutagenesis consists in a systematic alanine substitution. All side chain atoms past the β-carbon are subsequently removed. By comparing the relative binding energy between the wild type species and the alanine mutated, one can infer the contribution of each side chain. Alanine was chosen as it lacks unusual backbone dihedral angle preferences, contrary to glycine, for instance, that would in turn introduce conformational flexibility into the protein backbone. Performing alanine mutagenesis is a very time consuming technique, since each alanine-substituted protein must be separately constructed, expressed and refolded. Thus, Kollman and Massova[43] developed an approach for the in silico evaluation of the changes in the binding free energies as a result of mutating the residues of the interacting proteins. In their method MD simulation of protein-peptide complexes are performed, with the peptide residues being sequentially mutated to alanine. From this trajectory, data for the complex, the protein alone and the peptide alone are collected. The binding free energies are estimated as following:

$$\Delta G_{binding} = \Delta G_{wat}^{complex} - [\Delta G_{wat}^{protein} + \Delta G_{wat}^{peptide}] \qquad (29)$$

The prior equation is the result of a thermodynamic cycle for the MM-PBSA method.[44]

Each $\Delta G_{wat}$ is computed with the following set of equations:

$$\Delta G_{wat} = E_{gas} + \Delta G_{solv} - TS$$

$$G_{solv} = G_{PB} + G_{nonpolar}$$

$$E_{gas} = E_{int} + E_{elec} + E_{vdW}$$

$$E_{int} = E_{bond} + E_{angle} + E_{torsion} \qquad (30)$$

The result of the computational alanine scanning for each mutation is the difference in $\Delta G_{binding}$ between the wild type peptide and the mutated type:

$$\Delta\Delta G = \Delta G_{wt} - \Delta G_{mutant} \qquad (31)$$

Therefore a negative number corresponds to an unfavorable substitution that diminishes the binding of the peptide to the protein.

### 2.6.2 3D Database searching

In a 3D database search one wants to identify molecules that satisfy the chemical and geometric requirements of the receptor. As such, a 3D database contains information about the conformational properties of the molecules contained within it. It also enables the identification of lead series that are structurally different from the hit. In general, 3D search is performed depending on the information available about the receptor. If no 3D structure of the target macromolecule is available, it's still possible to derive a model called pharmacophore, that indicates the common features of the available active compounds.

A three-dimensional pharmacophore specifies the spatial relationships between pharmacophoric groups, such as hydrogen-bond donors and acceptors, positively and negatively charged groups, and hydrophobic moieties. These relations are often expressed as distances or distance ranges, but can also incorporate other geometric measures such as angles and planes.

Database searching usually works by exploring the conformational space for each molecule. and rejecting those that cannot satisfy the requirements of the pharmacophore. In addition, especially for searches of mimics of peptides, a shape similarity information could be added. Shape complementarity is a critical factor in molecular recognition between ligands and their receptors. Pharmacophore and shape technologies, if used separately, could in fact lead to false positives. As a consequence, new databases tend to incorporate both pharmacophoric and shape similarities. For example pepMMsMIMIC,[45] a web-oriented peptidomimetic compound virtual screening tool used in this thesis, suggests which chemical structures are able to mimic the protein-protein recognition of the 3D peptide bound to the protein using both pharmacophore and shape similarity. In Figure 3**Errore. L'origine riferimento non è stata trovata.** the pepMMsMIMIC algorithm is shown. Ref 45 contains further details.

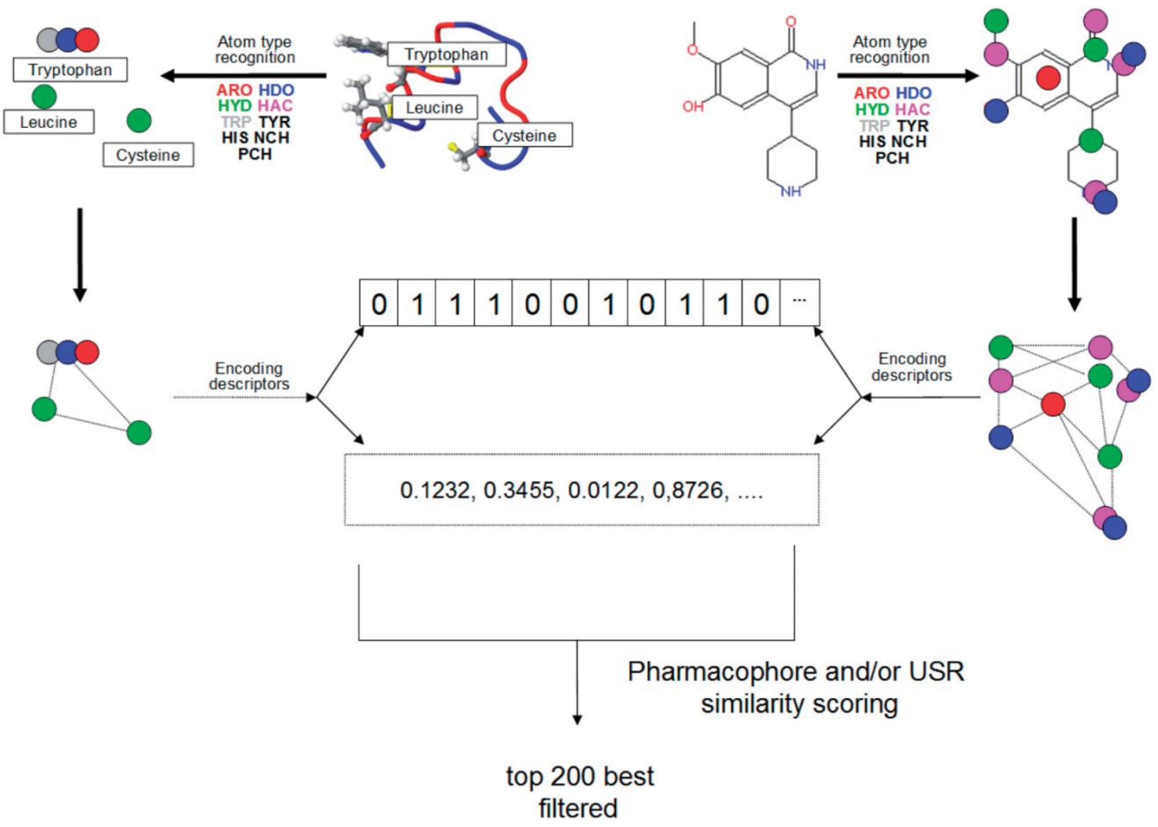


Figure 3. Workflow of pepMMsMIMIC.

### 2.6.3 Molecular docking

Molecular docking is mainly used in drug discovery to gain two valuable pieces of information:

- identify correct poses of a ligand in the binding site of the receptor
- estimate the strength of the ligand-receptor interaction

Because the synthesis and biological testing of leads is expensive and time consuming, suitable targets can be identified by docking in a reasonable time frame. For it to be working, the 3D structure of the receptor has to be available. It is worth mentioning at this point that the majority of the algorithms developed to position the ligand in the binding site only consider ligand flexibility, with the receptor treated as rigid, in order to reduce the computation time.

Different algorithms exist both for finding the best ligand geometry fitting the binding site and for estimating the strength of the binding and a full explanation can be found in various reviews.[46–48] The first problem has been tackled in order to avoid a full systematic search of all possible conformations of the ligand in the vicinity of the binding site. The methods can be summarized in the following categories:

- stochastic Monte Carlo methods, used by Glide,[49] make use of both random generation of conformations and energy minimization.
- fragment-based methods, used by FlexX, in which the ligand is first divided into fragments, each of them is docked into the active site and finally reconstructed in an incremental manner.[48]
- evolutionary-based methods, used by GOLD and AutoDock, use a genetic algorithm to generate the poses of the ligand inside the active site.
- shape complementarity methods, used by LigandFit, in which ligand conformations are accepted in the binding site if their shape fits that of the cavity.

As per measuring the strength of the binding, a number of different scoring functions have been developed. They all fall into three categories:

- force-field based methods, in which the score is obtained summing in intermolecular van der Waals and electrostatic interaction between the ligand and the receptor, together with a term related to the internal strain of the ligand and a term describing the desolvation energies of the two partners.
- empirical methods, which estimate the binding affinity of a complex on the basis of a set of weighted energy terms. The weighting coefficients are determined by fitting the binding affinity data of a training set with known three-dimensional structures.
- knowledge-based, or statistical methods, which employ potentials that are derived from the structural information of already available atomic structures.

In this thesis, the software Glide has been used to perform molecular docking. Glide uses a series of filters to search for possible locations of the ligand in the active site, and then to generate the best ligand binding poses through a coarse screening. The filter examines steric complementarity of the ligand to the protein and evaluates various ligand–protein interactions with the empirical Glide Score function.[50] Next, the ligand binding poses selected by the initial screening are minimized in situ with the OPLS force field. Finally the score is used to rank the resulting ligand binding poses.

## 2.7 BIBLIOGRAPHY

(1) Leach, A. R. *Molecular modelling: principles and applications*; Morris, C., Ed.; Prentice Hall, **2001**.

(2) Frenkel, D.; Smit, B. *Understanding molecular simulation: from algorithms to applications*; Acad. Pres.; **2002**.

(3) Born, M.; Oppenheimer, R. *Ann. Phys.* **1927**, *20*, 457–484.

(4) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(5) Hornak, V.; Abel, R.; Okur, A. *Proteins: Struct., Funct., Bioinf.* **2006**, *725*, 712–725.

(6) Brooks, B.; Bruccoleri, R. *J. Comput. Chem.* **1983**, *4*, 187–217.

(7) MacKerell,, a. D.; Bashford, D.; Dunbrack,, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(8) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(9) Allinger, N. L.; Kok, R. A.; Imam, M. R. *J. Comput. Chem.* **1988**, *9*, 591–595.

(10) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8566–8575.

(11) Allinger, N. L.; Chen, K.; Lii, J.-H. *J. Comput. Chem.* **1996**, *17*, 642–668.

(12) Halgren, T. A. *J. Comp. Chem.* **1996**, *17*, 616–641.

(13) Jorgensen, W. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(14) Hockney, R. .; Goel, S. .; Eastwood, J. . *J. Comput. Phys.* **1974**, *14*, 148–158.

(15) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(16) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(17) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144–1149.

(18) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(19) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(20) Saito, M. *J. Chem. Phys.* **1994**, *101*, 4055–4061.

(21) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(22) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(23) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(24) Morishita, T. *J. Chem. Phys.* **2000**, *113*, 2976–2982.

(25) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(26) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(27) Guarnieri, F.; Still, W. C. *J. Comput. Chem.* **1994**, *15*, 1302–1310.

(28) Saunders, M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. *J. Am. Chem. Soc.* **1990**, *112*, 1419–1427.

(29) Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

(30) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671–80.

(31) Kapral, R.; Hynes, J. T.; Ciccotti, G.; Carter, E. A. *Chem. Phys. Lett.* **1989**, *156*, 472–477.

(32) Bash, P.; Singh, U. C.; Langridge, R.; Kollman, P. *Science* **1987**, *236*, 564–568.

(33) Patey, G. N.; Valleau, J. P. *J. Chem. Phys.* **1975**, *63*, 2334–2339.

(34) Swendsen, R.; Ferrenberg, A. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.

(35) Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169–9174.

(36) Gullingsrud, J. R.; Braun, R.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 190–211.

(37) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.

(38) Bussi, G.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2006**, *96*, 090601.

(39) Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.

(40) Bonomi, M.; Parrinello, M. *Phys. Rev. Lett.* **2010**, *104*, 190601.

(41) Hodges, R. S.; Merrifield, R. B. *Int. J. Pept. Protein Res.* **1974**, *6*, 397–405.

(42) Fersht, A. R.; Shi, J.-P.; Knill-Jones, J.; Lowe, D. M.; Wilkinson, A. J.; Blow, D. M.; Brick, P.; Carter, P.; Waye, M. M. Y.; Winter, G. *Nature* **1985**, *314*, 235–238.

(43) Massova, I.; Kollman, P. a. *J. Am. Chem. Soc.* **1999**, *121*, 8133–8143.

(44) Kollman, P. a; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. a; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(45) Floris, M.; Masciocchi, J.; Fanton, M.; Moro, S. *Nucleic Acids Res.* **2011**, *39*, W261–269.

(46) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. *J. Med. Chem.* **2006**, *49*, 5851–5855.

(47) Moro, S.; Bacilieri, M.; Deflorian, F. *Expert Opin. Drug Discov.* **2007**, *2*, 37–49.

(48) Yuriev, E.; Agostino, M.; Ramsland, P. a *J. Mol. Recognit.* **2011**, *24*, 149–164.

(49) Glide, version 5.7, Schrödinger, LLC, New York, NY, **2011**.

(50) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput. Aided. Mol. Des.* **1997**, *11*, 425–445.

# CONFORMATIONAL STUDIES OF UNNATURAL GLYCOPEPTIDES

# Chapter 3: Glycopeptides and Glycoproteins

## 3.1 Introduction

Glycopeptides comprise a carbohydrate domain and a peptide domain. The carbohydrate can either be a single monosaccharide or a complex, potentially branched, oligosaccharide formed of up to about 20 monosaccharide units.

Glycoproteins, the larger version of glycopeptides, are crucial to many biological processes. An incomplete list include immune defense, viral infection, cell growth, cell-cell adhesion, and inflammation.[1,2] Glycoproteins are in fact the major constituents of the outer layer of mammalian cells and act as recognition elements.[3] Carbohydrate expression at the cell surface varies during the life cycle of the cell, because glycosyl transfer enzymes operate on it, especially during development.[4]

Carbohydrates, contrary to polypeptides and nucleic acids, can be highly branched, and their monomeric units can be attached to one another by many different linkage types. Different linkage positions (1→1, 2, 3, 4, 6 for hexopyranose), two anomeric configurations (α/β), change in ring size (furanose/pyranose) and the possibility of introducing site specific substitutions such as acetylation, phosphorylation or sulfation, induce an enormous structural diversity.[5]

Two major classes of glycosidic linkages to proteins exist (Figure 4): either they involve oxygen in the side chain of serine, threonine, or hydroxylysine (O-linked glycans) or nitrogen in the side chain of asparagine (N-linked glycans). In order to be glycosylated, asparagine has to be part of the triplet Asn-X-Ser, X being any amino acid but proline. In the case of O-glycosylation, a consensus sequence (Cys-XX-Gly-Gly-Ser/Thr-Cys) seem to correlate with O-fucosylation in epidermal growth factor domains.[6]

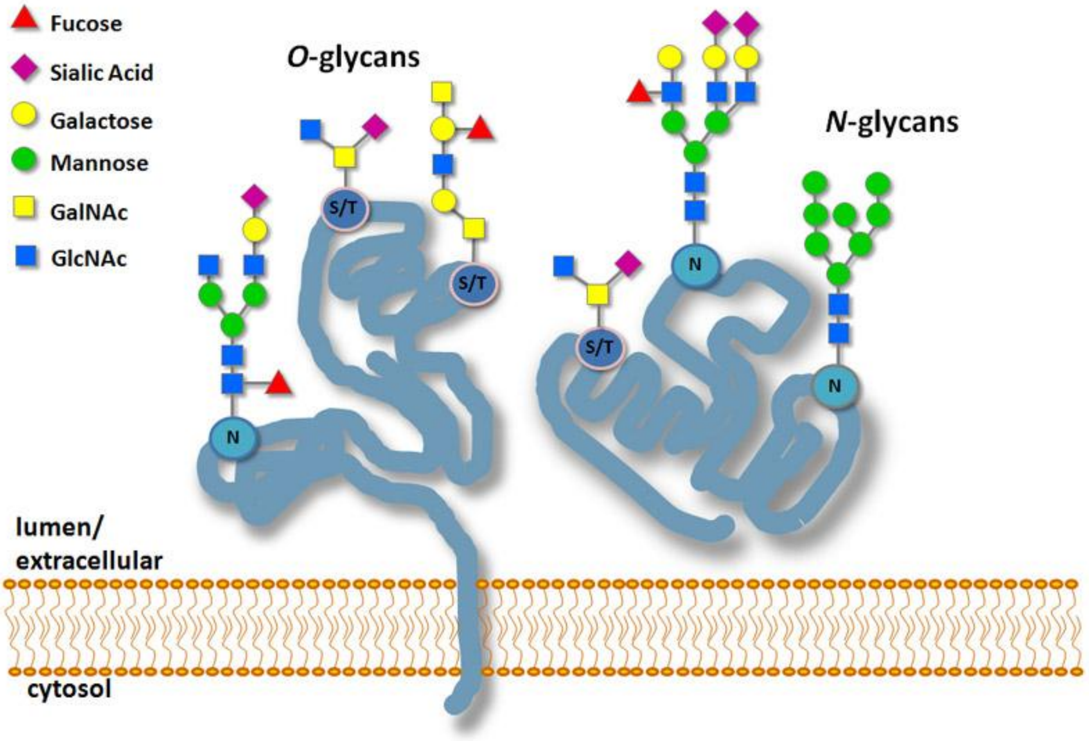


Figure 4. A graphic depiction of the two major forms of glycans.

All N-linked glycans comprise the pentasaccharide Manα1-6(Manα1-3)Manβ1-4GlcNAcβ1-4GlcNAc. On the contrary, O-linked glycans do not show any common core structure. Usually, glycans are linked to serine or threonine through GalNAc, although the linkage can also occur through fucose.[7]

Cell type determines the size and type of glycosylation, which is species and tissue specific.[8] The initial glycosylation reaction occurs biochemically with the transfer of a conserved tetradecasaccharide (GlcNAc2Man9Glc3) from the corresponding dolichyl-pyrophosphate-linked donor (Figure 5)

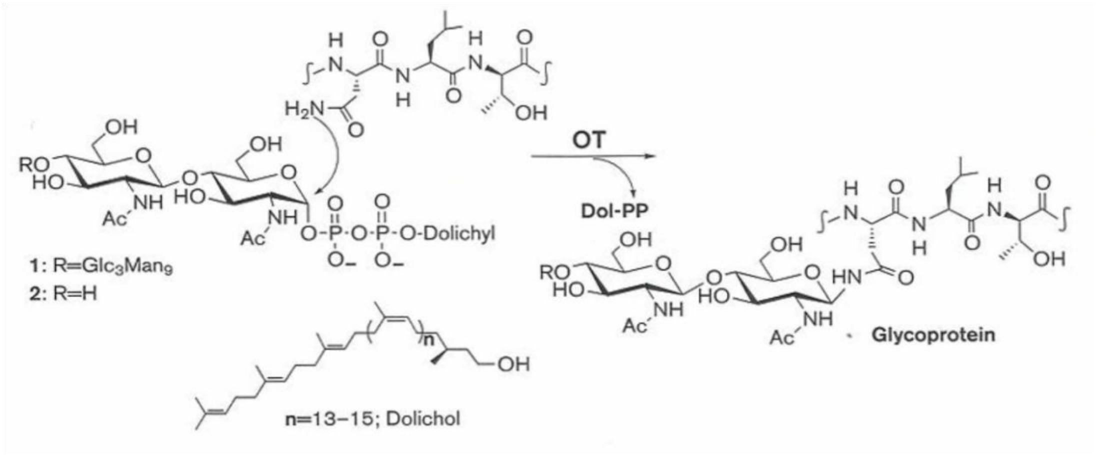


Figure 5. Asparagine-linked glycosylation.

Glycosylation takes place in the endoplasmic reticulum (ER) and the Golgi apparatus, where membrane-bound enzymes add monomeric carbohydrate units as the newly forming glycopeptide moves through the ER and Golgi apparatus. If an individual enzyme reaction does not go to completion, it gives rise to glycosylated variants of the polypeptide, called *glycoform*. Glycoproteins formation is influenced by physiological events, such as pregnancy, but also by diseases which have an effect on the enzymes in the cell.

### 3.2 GLYCOSYLATION'S EFFECT ON PROTEINS

Even though the importance of glycosylation is not yet fully understood, it has been proven that without glycosylation, immature proteins could misfold and be degraded before leaving the ER.[9] Essentially, glycosylation has an effect on the final folded conformation of newly generated polypeptides. Modifications in oligosaccharides displayed on cell surface are connected with various pathological effects. In fact, change in the population of glycoforms, which differ from the original glycoprotein for the position or the sequence of the sugar attached to the polypeptide, is caused by different diseases. The carbohydrate-deficient disorders, for instance, are a collection of rare diseases involving nervous-system disorders, which cause growth retardation, anomalous ocular movements, and infertility.[10,11] In this pathology, a variety of serum glycoproteins showed abnormal expression with respect to their glycoform populations. In addition, multiple organ dysfunctions found on patients have been correlated with 15 genes' mutations causing a deficient dolichol-linked oligosaccharide biosynthesis (Figure 5).

Variation in mucin expression and aberrant glycosylation are associated with cancer development.[12] Mucins are large extracellular glycoproteins and they act as a selective barrier at the epithelial cell surface.[13] The first involvement of mucins in cancer was the report of their high concentration levels in adenocarcinomas.[14] Immunohistochemical experiments have identified numerous Tumor-associated antigens (TAAs) on mucins.[15] Most TAAs on mucins were at first recognized as oligosaccharide structures, and many were identified as Blood group antigens; however, some of the antibodies were ultimately confirmed to recognize protein epitopes that were affected by glycosylation.[16,17] For this reason, antibodies against TAAs on mucins are used as diagnostic tools in cancer. The MUC1 mucin, for instance, is aberrantly glycosylated in cancer and can be detected in serum of late-stage patients.[18] Glycopeptides $T_N$ and the sialyl $T_N$ antigens are the two most important TAAs in MUC1 and are found in epithelial ovarian cancer, colon and breast which are found in human colon cancer, ovarian cancer and breast cancer (Figure 6).[19]

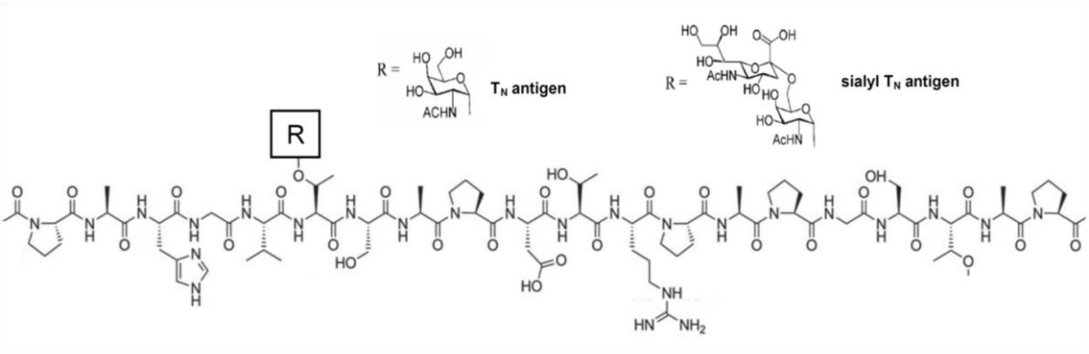


Figure 6. Tumour associated antigens $T_N$ and sialyl $T_N$.

Furthermore, overexpressed mucins MUC1 can produce autoantibodies that might serve as diagnostic biomarkers for cancer-diagnosis[20] but also for the detection of Multiple Sclerosis (MS), the most recurrent chronic inflammatory disease of the central nervous system and the most common source of disability in adults.[21] Papini et al.[22] reported in fact that abnormal N-glucosylation triggered an autoantibody response in MS and observed for the first time autoantibodies in MS patients' sera.

Human immunodeficiency (HIV) infection is a classic example of the role of glycoproteins in viral infection. The HIV type 1 (HIV-1) envelope glycoprotein comprises two non-covalently associated subunits, gp120 and gp41, which result from proteolytic cleavage of a precursor polypeptide, gp160. Gp120 in particular is responsible for the target-cell recognition through interaction with the cell-surface receptor CD4.[23]

Despite massive scientific effort, the development of a vaccine against HIV has not yet successfully accomplished. Commonly utilized vaccine formulations have been unable to elicit potent and broadly neutralizing immune responses,[24] due to the high rate of viral mutations. Another serious obstacle concerns that fact that the protein domain of the viral surface envelope protein gp120 becomes extensively glycosylated and shows very low immunogenicity.[25] In fact, huge heterogeneity among individual isolated HIV-1 is observed in patients. Gp120 is typically modified with carbohydrates in order to protect the polypeptide domain from recognition by the immune system.[26] These glycans could then be themselves the targets for an anti-HIV vaccine. Interestingly, it has been reported that 2G12, a potent anti-HIV antibody, seem to bind to the hybrid- or high-mannose type carbohydrate domains of gp120 (Figure 7).[27]

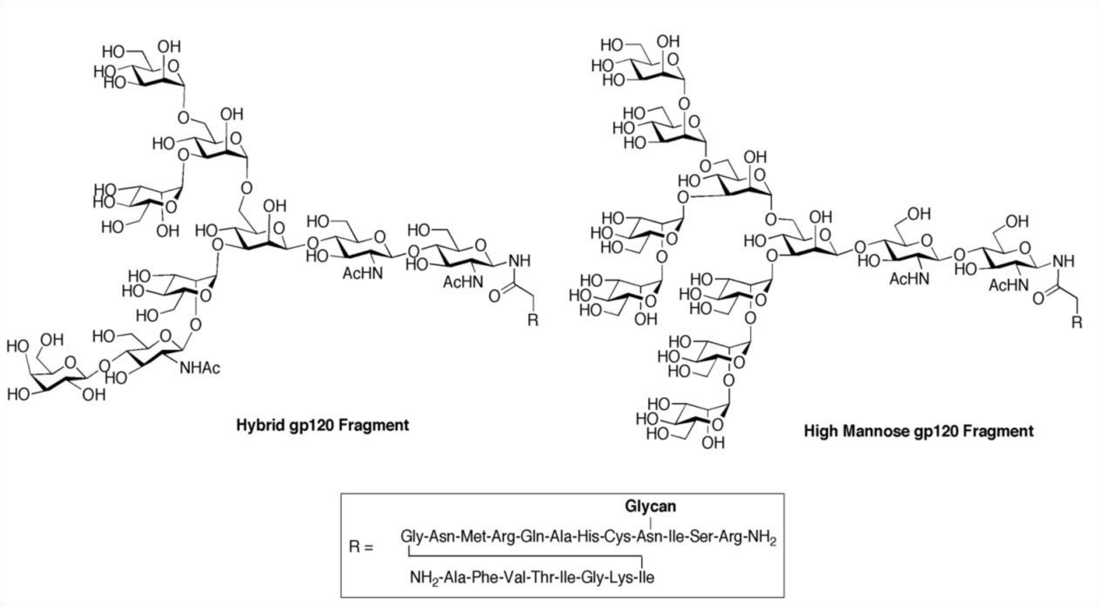

Hybrid gp120 Fragment
High Mannose gp120 Fragment
Glycan
R = Gly-Asn-Met-Arg-Gln-Ala-His-Cys-Asn-Ile-Ser-Arg-NH2
NH2-Ala-Phe-Val-Thr-Ile-Gly-Lys-Ile

Figure 7. Hybrid and high-mannose gp120 fragments.

Identifying antigens that resemble these natural epitopes is an important step toward the development of HIV-1 vaccines.[26]

Relevance of glycopeptides is highlighted also in the design of a GPI-based anti-toxic malaria vaccines. GPI anchor is a glycolipid that can bind to the C-terminus of a protein during post-translational modification. It comprises a glycan core bridged to the C-terminal amino acid of a protein via an ethanolamine phosphate. A lipid chain anchors the protein to the cell membrane (Figure 8).

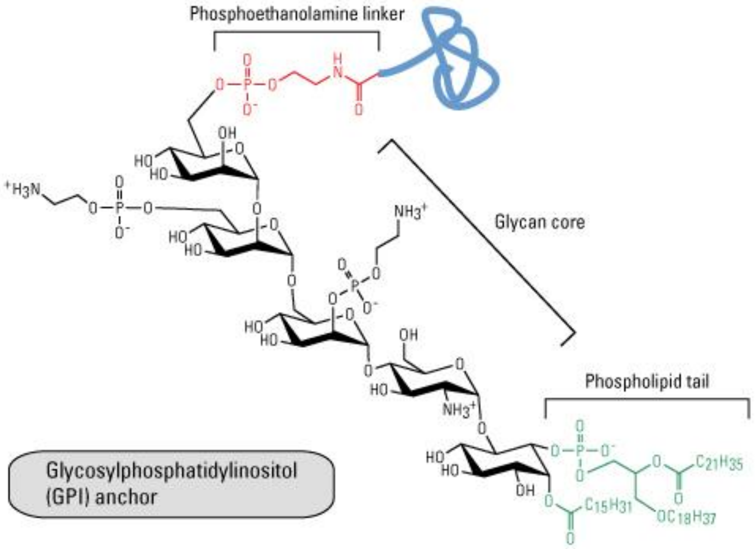

Phosphoethanolamine linker
Glycan core
Phospholipid tail
Glycosylphosphatidylinositol (GPI) anchor

Figure 8. GPI-anchored protein.

GPIs are conserved glycolipids found in the outer cell membranes of mostly all eukaryotic cells and make up for the 90% of protein glycosylation in protozoan parasites.[28,29] In fact proteins are often modified after being translated by glycosylation and lipidation.[30] GPI anchors include both types of modification and link many proteins to the surface of the cell. The malaria parasite Plasmodium falciparum shows on its cell surface the GPI glycolipid. This highly conserved endotoxin may contribute to pathogenesis in humans. A recent study in which a non-toxic GPI oligosaccharide was coupled to a carrier protein, produced an immune response and enabled enhanced protection against malaria in a preclinical trial.[31] Synthetic GPI, from Seeberger group,[32] is therefore a prototype carbohydrate anti-toxic vaccine against malaria.

Since glycoproteins have been extensively used as therapeutics in modern medicine,[33] during the years has emerged a growing interest in understanding the impact of glycosylation on protein folding. So, various experimental, computational, and bioinformatic experiments were performed to define glycosylation-induced effects on protein structure.[34–37] Glycosylation seems to strengthen the stability of the folded protein via long range H-bonds and hydrophobic interactions between the sugar moieties and the protein.[38] Various observations highlighted the importance of asparagine glycosylation for the proper folding and assembly in the biosynthesis of proteins: indeed, presence of N-linked glycosylation inhibitors, such as tunicamycin, often brings an incomplete or incorrect folding.[39,40]

The 3D arrangement of the individual protein determines the type and extent of its glycosylation. A number of factors may be involved, such as:

a) The position of the glycosylation sites in the protein. N-Linked sites at the exposed turns of β-sheets, are usually in use while those close to the C-terminus are more often unoccupied.
b) The interaction of the emergent oligosaccharide with the protein surface. This may affect the glycan conformation by modifying its accessibility to specific glycosylenzymes.
c) The interaction of protein subunits to form oligomers. This could preclude glycosylation or limit the glycoforms at specific sites.

Five glycoproteins were subjected by Giartosio and coworkers to enzymatic deglycosylation with different glycosidases in order to obtain deglycosylated products.[41] Although circular dichroism (CD) measurements suggest that secondary structure motives were not disrupted by glycosylation, comparison of the unfolding temperatures suggested improved stability in the glycosylated proteins.

Carbohydrates affect other physicochemical properties of proteins: in arctic fish,[42] O-glycan-rich proteins act as natural “antifreeze”, preventing nucleation of ice and permitting life at low temperatures (Chapter 4).

## 3.3 STRUCTURAL ANALYSIS OF GLYCOPEPTIDES

Various techniques have been used to analyze glycopeptide and glycoprotein structures. Circular dichroism (CD) and Nuclear magnetic resonance (NMR) spectroscopies, molecular-modelling techniques, fluorescence Resonance Energy Transfer (FRET) and site-directed mutagenesis studies have all permitted a better understanding of the effect of oligosaccharide attachment on the structure and stability of the peptide chain. X-ray crystallography is more difficult to apply to glycoproteins, because of the heterogeneity of glycan structures and the conformational flexibility of the saccharide moieties on the protein surface. Detailed analysis of the glycosylation effect is hampered due to the fact that glycoproteins are large, as they usually contain multiple subunits. Hence, the study of large glycoproteins at an atomistic level, which could offer insight on the site-specific effects of glycosylation, is still hardly feasible. Furthermore, biophysical studies of natively glycosylated proteins are prevented by the scarce availability of defined materials for the experiments, which is mostly due to the intrinsic heterogeneity of glycoproteins. As a consequence, first studies were directed to the conformational analysis of defined glycopeptides. Kahne et al. in 1993[43] investigated the conformation of the backbone of a linear hexapeptide (Phe-Phe-D-Trp-Lys-Thr-Phe) being glycosylated. The results showed that glycosylation with a single monosaccharide (GalNAc) has a deep effect on the backbone conformation, restraining the conformational space available to the peptide and favoring conformations in which the peptide chain bends away from the GalNac moiety. In a subsequent study, a disaccharide unit was attached to the same peptide[44]. NMR allowed the evaluation of the differences in the backbone conformation between the glycosylated and the non glycosylated peptides. NOEs between sequential amide protons are used as indicator of the average backbone conformation, and they were used to point out that the attachment of a second monosaccharide changed the backbone conformation with respect to the monoglycosylated hexapeptide. The explanation for the observed changes provided by the authors was the exclusion of many conformations for steric reasons. Restriction of the conformational space was also involved, in the authors' opinion, in glycoprotein folding and thermal stability.

Early works on the effect of glycosylation on the conformational mobility of oligopeptides[45] indicated that glycosylation could change the conformational profile of a polypeptide and enable the sampling of conformational space originally forbidden to it. The same conclusions were also obtained in more recent works.[46] Powers et al. investigated the folding process of the mono-N-glycosylated adhesion domain of the human immune cell receptor cluster of differentiation 2 (hCD2ad, Figure 9) and systematically examined the influence of the N-glycan on the folding energy profile.[34] hCD2ad, a representative of the immunoglobulin (Ig) superfamily, is a small glycoprotein (105 residues) with many β-strands secondary structure elements. N-glycan structures accelerate folding by 4-fold and stabilize the structure by 3.1 kcal/mol, relative to the non-glycosylated protein. The N-glycan's first saccharide unit is responsible for the entire increase of folding rate and for 2/3 of the native state stabilization. The remaining third of the stabilization is due from the successive two saccharide units. Thus, the conserved N-linked triose core, $ManGlcNAc_2$, speeds up both the kinetics and the thermodynamics of protein folding.

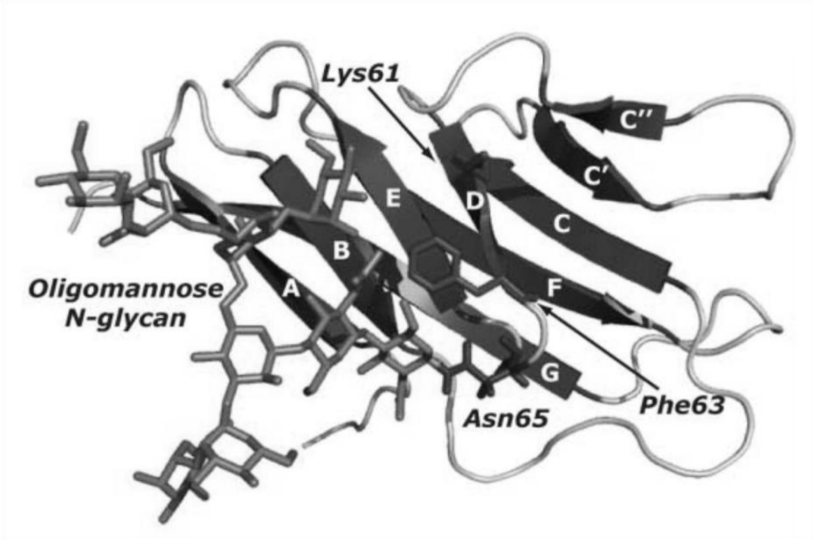


Figure 9. Structure of hCD2ad with N-glycan.

Recently, the effect of glycosylation on protein folding was explored by computational methods by Levy and coworkers.[35] The folding of the SH3 domain was simulated using a native topology-based (Go) model. The SH3 domain is a small protein (56 amino acids, Figure 10) whose folding is well characterized, both experimentally and theoretically. In this case, SH3 domain has been glycosylated with different numbers of polysaccharides at different sites on the protein's surface. Although the SH3 domain is not a glycoprotein, studying a protein whose folding is well known could give an insight into the common effects of glycosylation on folding. The authors found that thermodynamic stabilization correlated with the degree of glycosylation and, to a lesser extent, with the size of the polysaccharides. The stabilization effect depended upon the position of the glycans; thus, the same degree of glycosylation could produce different thermal effects, depending on the location of the sugars. This study suggests that glycosylation can alter the biophysical properties of proteins and offers a new way to design thermally stabilized proteins.

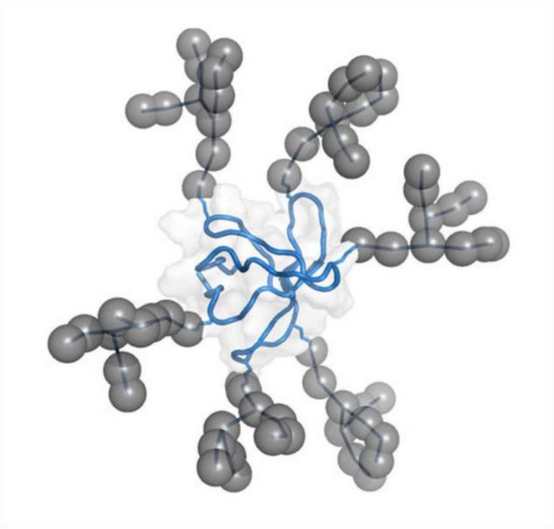


Figure 10. SH3 domain: the polypeptide chain is in blue and the carbohydrate rings are the gray balls. Each glycan contains 11 sugars.

The local structure and the stability of small glycopeptides were also extensively studied in the group of Imperiali.[47] After the synthesis of short glycopeptides in which key molecular elements of the sugar, particularly the N-acetyl groups, were modulated, they explored the effect of variations in carbohydrate composition on the glycopeptide backbone conformation. The short oligopeptide AcNH-Orn-Ile-Thr-Pro-Asn-Gly-Thr-Trp-Ala-CONH2, based on the glycosylation site of the hemagglutinin protein of influenza virus, was synthesized and derivatized with five different carbohydrates. The different glycopeptides conformations were then verified using 2D NMR methods. The nonglycosylated peptide preferred conformation was found to be an Asx-turn,[48] with an H-bond forming between the asparagine side chain and the peptide backbone (Figure 11a) β-Chitobiose, on the Asn side chain, induces a native β-turn structure (Figure 11b). The addition of a large substituent to this key amino acid could produce a prevalence of the β-turn over the Asx turn. The Asx turn could be disfavored in the glycosylated state for steric reasons.

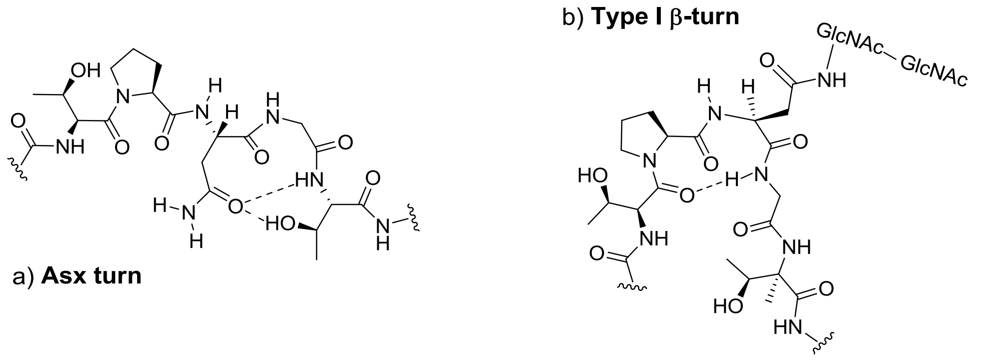


Figure 11. Comparison of the Asx turn and β-turn conformations in a Thr-Pro-An-Gly-Thr-sequence. The Asx turn is formed by the Asn side chain.

Neo-glycoconjugates, including unnatural carbohydrate moieties, were used to assess the role of the sugar in stabilizing a β-turn conformation: glycosylations, using β-N-linked GlcNAc-GlcNAc, Glc-GlcNAc, GlcNAc, GlcNAc-Glc and Glc-Glc, have highlighted the critical role of the N-acetyl group of the proximal sugar for inducing a β-turn peptide conformation. It is also worth noting that the glycopeptide derivatized with Gal-GlcNAc failed to generate a β-turn conformation, and instead was found to adopt an extended structure, suggesting a highly specific carbohydrate conformation.

Afterwards, Imperiali et al.[49] performed a comparison between an α- and a β-linked glycopeptide and the corresponding non glycosylated peptide (Figure 12). 2D-NMR experiments were used to assess the conformational properties of both the new α-linked glycopeptide and the unglycosylated peptide, as well as the β-linked glycopeptide. From the NMR results the authors derived that the stereochemistry at the anomeric center of the N-linked carbohydrate has a dramatic effect on the conformation of the peptide backbone. Indeed, only the β-linked glycopeptide is in a stable β-turn conformation.

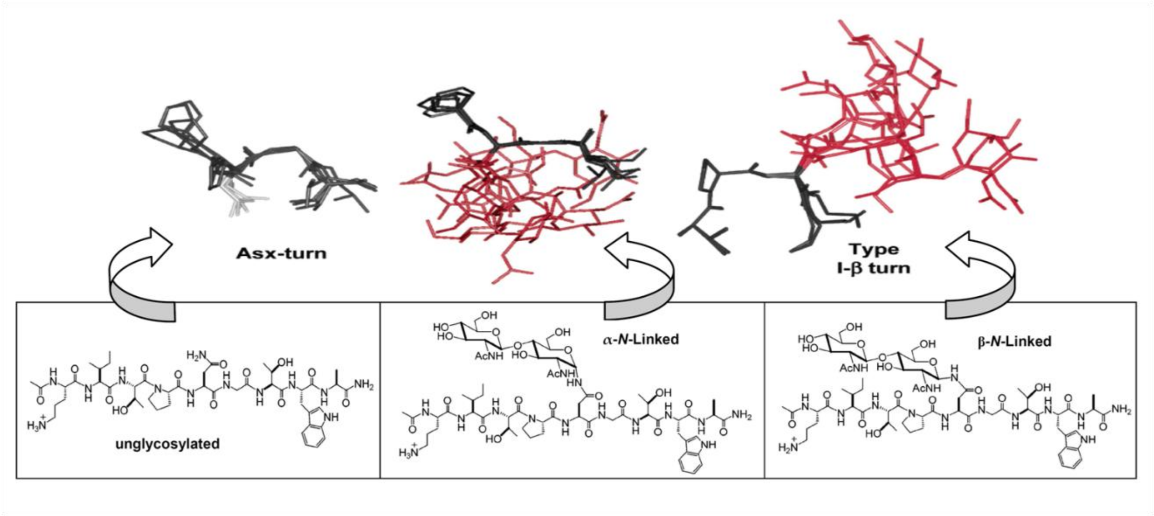


Figure 12. Comparison of the 3D structures between non-glycosylated peptide, α-linked glycopeptide, and β-linked glycopeptide.

The α-N-linked glycopeptide mainly adopted a conformation similar to that of the non-glycosylated peptide, an Asx-turn structure. Corresponding computational modelling of these glycopeptides, via explicit solvent Molecular Dynamics simulations, obtained the same results and independently predicted the NMR experiments.

To our knowledge, this is the only study that assessed the different conformation of non natural α-N-linked glycopeptide, with respect to non-glycosylated peptides and from the natural β-*N*-linked glycopeptide, suggesting the possibility that α-N-linked glycopeptides could display new structural properties.

## 3.4 BIBLIOGRAPHY

(1) Bertozzi, C. R. *Science* **2001**, *291*, 2357–2364.

(2) Rudd, P. M.; Elliott, T.; Cresswell, P.; Wilson, I. A.; Dwek, R. A. *Sci.* **2001**, *291* , 2370–2376.

(3) Taylor, C. M. *Tetrahedron* **1998**, *54*, 11317–11362.

(4) Varki, A. *Glycobiol.* **1993**, *3* , 97–130.

(5) Gamblin, D. P.; Scanlan, E. M.; Davis, B. G. *Chem. Rev.* **2008**, *109*, 131–163.

(6) Harris, R. J.; van Halbeek, H.; Glushka, J.; Basa, L. J.; Ling, V. T.; Smith, K. J.; Spellman, M. W. *Biochemistry* **1993**, *32*, 6539–6547.

(7) Fukuda, M.; Hindsgaul, O. *Molecular glycobiology*; IRL Press, 1994.

(8) Parekh, R. B.; Dwek, R. A.; Thomas, J. R.; Opdenakker, G.; Rademacher, T. W.; Wittwer, A. J.; Howard, S. C.; Nelson, R.; Siegel, N. R.; Jennings, M.; Harakas, N.; Feder, J. *Biochemistry* **1989**, *28*, 7644–7662.

(9) Wyss, D. F.; Wagner, G. *Curr. Opin. Biotechnol.* **1996**, *7*, 409–416.

(10) Jaeken, J.; van Eijk, H. G.; van der Heul, C.; Corbeel, L.; Eeckels, R.; Eggermont, E. *Clin. Chim. Acta* **1984**, *144*, 245–247.

(11) Jensen, H.; Kjaergaard, S.; Klie, F.; Moller, H. U. *Ophthalmic Genet.* **2003**, *24*, 81–88.

(12) Sørensen, A. L.; Reis, C. A.; Tarp, M. A.; Mandel, U.; Ramachandran, K.; Sankaranarayanan, V.; Schwientek, T.; Graham, R.; Taylor-Papadimitriou, J.; Hollingsworth, M. A.; Burchell, J.; Clausen, H. *Glycobiology* **2006**, *16*, 96–107.

(13) Hollingsworth, M. A.; Swanson, B. J. *Nat. Rev. Cancer* **2004**, *4*, 45–60.

(14) Hukill, P. B.; Vidone, R. A. *Lab. Invest.* **1965**, *14*, 1624–16235.

(15) Goldenberg, D. M.; Pegram, C. A.; Vazquez, J. J. *J. Immunol.* **1975**, *114*, 1008–1013.

(16) Bramwell, M. E.; Bhavanandan, V. P.; Wiseman, G.; Harris, H. *Br. J. Cancer* **1983**, *48*, 177–183.

(17) Burchell, J.; Durbin, H.; Taylor-Papadimitriou, J. *J. Immunol.* **1983**, *131*, 508–513.

(18) Rughetti, A.; Pellicciotta, I.; Biffoni, M.; Bäckström, M.; Link, T.; Bennet, E. P.; Clausen, H.; Noll, T.; Hansson, G. C.; Burchell, J. M.; Frati, L.; Taylor-Papadimitriou, J.; Nuti, M. *J. Immunol.* **2005**, *174*, 7764–7772.

(19) Tashiro, Y.; Yonezawa, S.; Kim, Y. S.; Sato, E. *Hum. Pathol.* **1994**, *25*, 364–372.

(20) Lu, H.; Goodell, V.; Disis, M. L. *J. Proteome Res.* **2008**, *7*, 1388–1394.

(21) Whetten-Goldstein, K.; Sloan, F. A.; Goldstein, L. B.; Kulas, E. D. *Mult. Scler.* **1998**, *4*, 419–425.

(22) Lolli, F.; Mulinacci, B.; Carotenuto, A.; Bonetti, B.; Sabatino, G.; Mazzanti, B.; D'Ursi, A. M.; Novellino, E.; Pazzagli, M.; Lovato, L.; Alcaro, M. C.; Peroni, E.; Pozo-Carrero, M. C.; Nuti, F.; Battistini, L.; Borsellino, G.; Chelli, M.; Rovero, P.; Papini, A. M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10273–10278.

(23) Chan, D. C.; Fass, D.; Berger, J. M.; Kim, P. S. *Cell* **1997**, *89*, 263–273.

(24) McMichael, A. J.; Hanke, T. *Nat. Med.* **2003**, *9*, 874–880.

(25) Wei, X.; Decker, J. M.; Wang, S.; Hui, H.; Kappes, J. C.; Wu, X.; Salazar-Gonzalez, J. F.; Salazar, M. G.; Kilby, J. M.; Saag, M. S.; Komarova, N. L.; Nowak, M. A.; Hahn, B. H.; Kwong, P. D.; Shaw, G. M. *Nature* **2003**, *422*, 307–312.

(26) Doores, K. J.; Fulton, Z.; Hong, V.; Patel, M. K.; Scanlan, C. N.; Wormald, M. R.; Finn, M. G.; Burton, D. R.; Wilson, I. A.; Davis, B. G. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 17107–17112.

(27) Sanders, R. W.; Venturi, M.; Schiffner, L.; Kalyanaraman, R.; Katinger, H.; Lloyd, K. O.; Kwong, P. D.; Moore, J. P. *J. Virol.* **2002**, *76*, 7293–7305.

(28) Ferguson, M. A. J.; Williams, A. F. *Annu. Rev. Biochem.* **1988**, *57*, 285–320.

(29) Gowda, D. C.; Davidson, E. A. *Parasitol. Today* **1999**, *15*, 147–152.

(30) Walsh, C. T.; Garneau-Tsodikova, S.; Gatto, G. J. *Angew. Chem. Int. Ed. Engl.* **2005**, *44*, 7342–7372.

(31) Schofield, L.; Hewitt, M. C.; Evans, K.; Siomos, M.-A.; Seeberger, P. H. *Nature* **2002**, *418*, 785–789.

(32) Becker, C. F. W.; Liu, X.; Olschewski, D.; Castelli, R.; Seidel, R.; Seeberger, P. H. *Angew. Chem. Int. Ed. Engl.* **2008**, *47*, 8215–8219.

(33) Solá, R. J.; Griebenow, K. *J. Pharm. Sci.* **2009**, *98*, 1223–1245.

(34) Hanson, S. R.; Culyba, E. K.; Hsu, T.-L.; Wong, C.-H.; Kelly, J. W.; Powers, E. T. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 3131–3136.

(35) Shental-Bechor, D.; Levy, Y. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 8256–8261.

(36) Petrescu, A.-J.; Milac, A.-L.; Petrescu, S. M.; Dwek, R. A.; Wormald, M. R. *Glycobiology* **2004**, *14*, 103–114.

(37) Shental-Bechor, D.; Levy, Y. *Curr. Opin. Struct. Biol.* **2009**, *19*, 524–533.

(38) O'Connor, S. E.; Imperiali, B. *Chem. Biol.* **1996**, *3*, 803–812.

(39) Riederer, M. A.; Hinnen, A. *J. Bacteriol.* **1991**, *173*, 3539–3546.

(40) Marquardt, T.; Helenius, A. *J. Cell Biol.* **1992**, *117*, 505–513.

(41) Wang, C.; Eufemi, M.; Turano, C.; Giartosio, A. *Biochemistry* **1996**, *35*, 7299–7307.

(42) Hansen, T. N.; Carpenter, J. F. *Biophys. J.* **1993**, *64*, 1843–1850.

(43) Hamilton Andreotti, A.; Kahne, D. *J. Am. Chem. Soc.* **1993**, *115*, 3352–3353.

(44) Liang, R.; Andreotti, A. H.; Kahne, D. *J. Am. Chem. Soc.* **1995**, *117*, 10395–10396.

(45) Matthews, C. R. *Annu. Rev. Biochem.* **1993**, *62*, 653–683.

(46) Wormald, M. R.; Dwek, R. A. *Structure* **1999**, *7*, R155–160.

(47) O'Conner, S. E.; Imperiali, B. *Chem. Biol.* **1998**, *5*, 427–437.

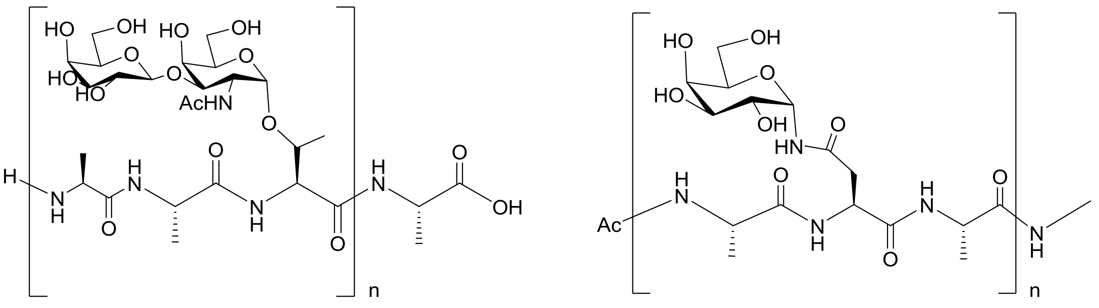
(48) O'Connor, S. E.; Imperiali, B. *J. Am. Chem. Soc.* **1997**, *119*, 2295–2296.

(49) Bosques, C. J.; Tschampel, S. M.; Woods, R. J.; Imperiali, B. *J. Am. Chem. Soc.* **2004**, *126*, 8421–8425.

# Chapter 4: Conformational analyses of α-*N*-linked glycopeptides

## 4.1 INTRODUCTION

Among natural glycopeptides, so-called antifreeze glycoproteins (AFGPs) are mucin-like glycopeptides, composed of a repeating unit (Ala-Thr-Ala) in which the disaccharide β-D-Gal-(1→3)-α-DGalNAc is bound to the threonyl residue (Figure 13, **101**).



**101**

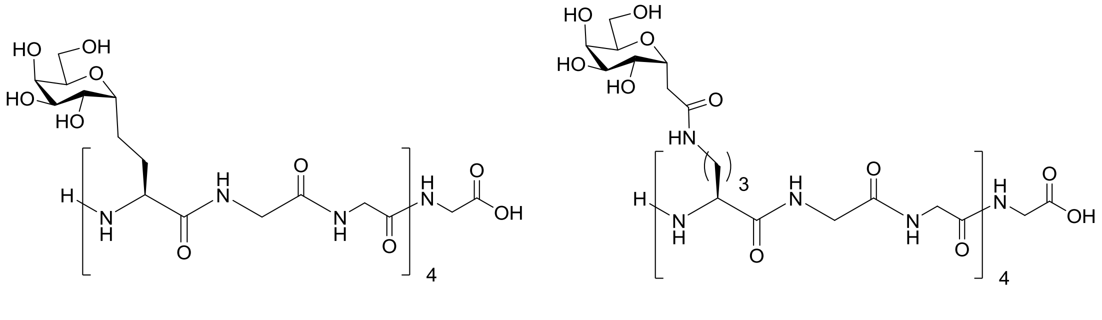
natural antifreeze glycopeptides (AFGP)

**102**

unnatural α-N-linked mimics of AFGP

Figure 13. Unnatural α-*N*-linked glycopeptides **102**, described in this Chapter, in comparison to natural AFGPs **101**.

Their relative molecular mass ranges from about 2000 to 33000 ($4 \leq n \leq 55$).[1] AFPs are found in the blood serum of fishes living in the sub-zero Arctic and Antarctic oceans. Indeed, polar fish could not survive without the presence of AFPs, since the main role of AFPs is to prevent ice crystal growth (IRI, ice recrystallization inhibition)[2] and to reduce the blood freezing point, thus creating a hysteresis between the melting and freezing points (TH, thermal hysteresis) of water.[3] The ability of inhibiting ice crystal growth makes AFPs useful in many areas of agriculture and in the frozen food industry.[4] Mechanistic studies performed on AFGPs are scarce, due to the lack of access to pure samples from natural sources, and the very limited quantities provided by unnatural sources. In fact, the first synthesis of AFGPs as pure glycoforms was not reported until 2004.[5] Here, also tentative structure-activity studies were performed. These studies showed a decisive role of the hydrophobic interactions, N-acetyl group and Ala-Thr-Ala backbone, in enabling the antifreeze activity. In particular TH was found to correlate to the number of repeating units of the molecules, reaching an optimal value for n = 5, 6, 7. These data were crucial to gain more insight into the mechanism of action of AFPs and to guide the design of AFP mimics. Recently, it has been proposed that the ability to selectively lower the freezing point of a solution could also help the preservation and hypothermic storage of biomedical supplies. So, the use of AFPs as cryoprotectants has also been explored.[6,7] Unfortunately, cells tend to damage if a temperature below the TH gap is reached.[8–10] However, as AFGPs, also promote the inhibition of crystal growth during ice recrystallization (IRI), they could be used to protect the cells from damage during the cryopreservation. AFGP mimics not possessing thermal hysteresis properties, but still able to inhibit ice recrystallization, have been recently published by Ben's group.[11–13] Two unnatural C-linked galactosyl AFGPs **103a** and **103b** (Figure 14) were synthesized. With respect to the natural counterpart, these molecules replace the alanine residue with glycine. They turned out to be strong inhibitors of ice recrystallization and were also found to shield embryonic liver cells from ice crystal damage at millimolar (mM) concentration. In addition, **103a** showed negligible in vitro cytotoxicity.



**103a**

**103b**

Figure 14. Unnatural C-linked glycopeptides **103** described by Ben as cryoprotectants.

The design and synthesis of glycopeptides is of great challenge, and could also help predict the antifreeze activity of new molecules. Hence, the synthesis of antifreeze mimics with general formula **102** (Figure 13) has been performed in Bernardi group[14] introducing the following modifications to natural AFGP repeating unit **101**:

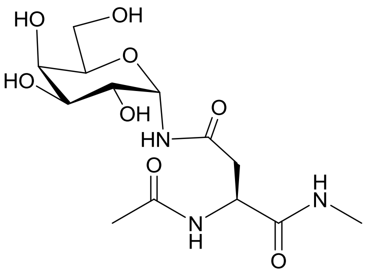
a) the repeating unit is the tripeptide Ala-Asn-Ala, with a similar hydrophobicity with respect to the natural AFGPs **101**.

b) A galactose moiety is used instead of the Gal-GalNAc disaccharide, following the C-linked unnatural glycopeptides **103a-b**.

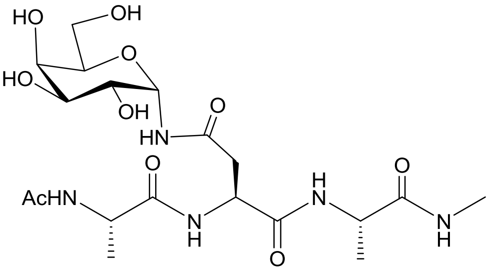
c) An α-*N*-linked galactosyl asparagine in place of the Gal-GalNAc-threonyl residues of natural AFGPs **101**.

To the best of our knowledge, the most recent conformational study of an unnatural α-*N*-linked glycopeptide was reported in 2004 by Imperiali and Woods.[15] The α-*N*-linked glycopeptide was found to have a conformation similar to the unglycosylated peptide and different from the β-*N*-linked glycopeptide. The peptide conformation of *N*-linked glycopeptides was hence found to depend on the anomeric configuration of the appended glycan.

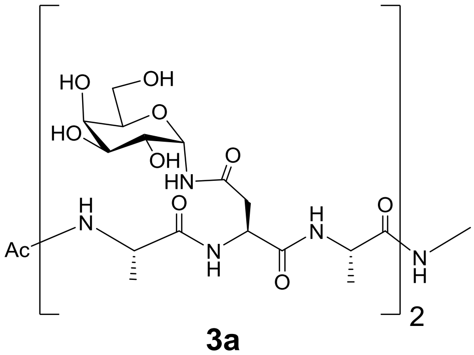
In this chapter the results of the conformational studies of a series of compounds having the general formula **101** will be presented, together with the complementary experimental characterizations of these molecules. In particular, the following molecules will be discussed (Figure 15):



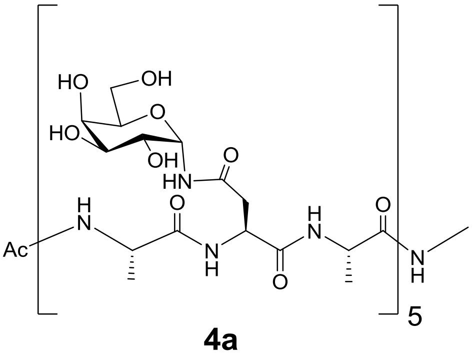
**1a**



**2a**



**3a**



**4a**

Figure 15. α-*N*-linked glycopeptides examined in this chapter.

Some of these compounds have also shown affinity towards the plant toxin *Viscum album agglutinin* (VAA), a model for lectin drug design[16] and the *Erythrina cristagalli* agglutinin (from coral tree, ECA).[17] The findings highlighted here have been published in 2012[c] in the *Organic & Biomolecular Chemistry* Journal.[18]

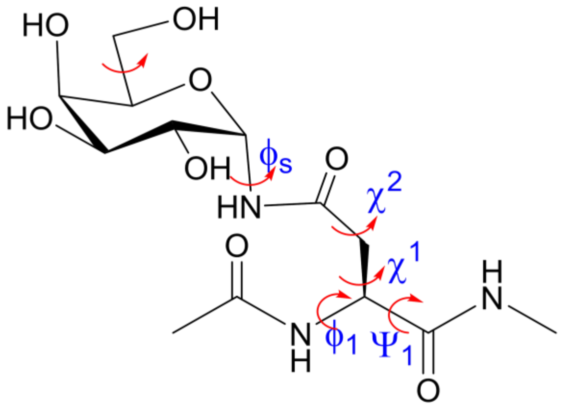
## 4.2 Conformational analyses of 1a and 2a

Compounds **1a** and **2a** were studied using a MC/EM method[19] followed by a MC/SD mix simulation (Chapter 2).[20] The MacroModel software[21], from Schrödinger, was used.

[c] Marcelo, F.; Cañada, F. J.; André, S.; Colombo, C.; Doro, F.; Gabius, H.-J.; Bernardi, A.; Jiménez-Barbero, J. *Org. Biomol. Chem.* **2012**, *10*, 5916–23.

### 4.2.1 Conformational search: 1a

A total number of 6 rotatable bonds were varied during the MC/EM calculation (Figure 16).



| | | |
|---|---|---|
| $\phi_s$ | → | O5s-C1s-N-$C^\gamma$ |
| $\chi^1$ | → | N-$C^\alpha$-$C^\beta$-$C^\gamma$ |
| $\chi^2$ | → | $C^\alpha$-$C^\beta$-$C^\gamma$-N |
| $\phi_1$ | → | C′-N-$C^\alpha$-C′ |
| $\psi_2$ | → | N-$C^\alpha$-C′-N |

Figure 16. Rotatable bonds varied during the conformational search for compound **1a**.

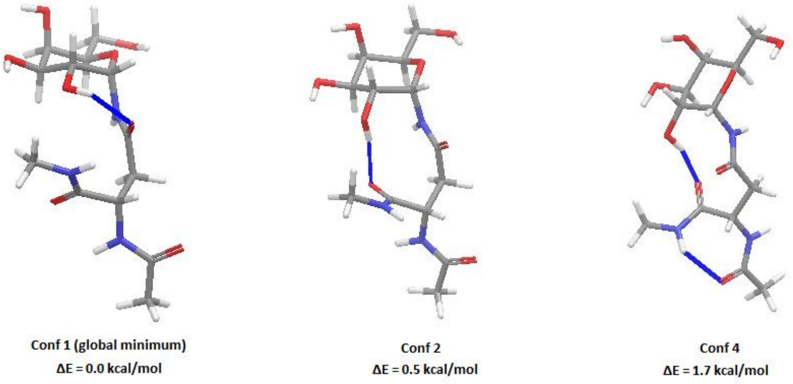
Following prior literature,[22] the dihedral angle referring to the anomeric torsion has been named $\varphi_S$. $\chi^1$ and $\chi^2$ are the dihedral angles of the Asn side chain.

A total number of 41 unique conformers were found in 5 kcal/mol from the global minimum, of which 2 were in the first kcal/mol and 5 in the first 2 kcal/mol.

| Conf. | $\Delta$E | $\psi_1$ | $\phi_1$ | $\chi^1$ | $\chi^2$ | $\phi_s$ | Boltz Pop | H-bonds |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 135.1 | -162.0 | 173.7 | -139.5 | 154.7 | 55.1 | 1 |
| 2 | 0.5 | 107.0 | -157.9 | 163.1 | -123.4 | 134.0 | 24.3 | 1 |
| 3 | 1.5 | 135.0 | -162.0 | 173.7 | -139.6 | 154.7 | 4.2 | 1 |
| 4 | 1.7 | 79.7 | -79.0 | 172.7 | -128.4 | 140.1 | 3.4 | 2 |
| 5 | 1.9 | 139.0 | -162.9 | 169.6 | 73.6 | 153.7 | 2.2 | 1 |
| 6 | 2.0 | 107.1 | -157.9 | 163.1 | -123.3 | 133.8 | 1.9 | 1 |
| 7 | 2.2 | 135.1 | -162.0 | 173.7 | -139.8 | 154.7 | 1.4 | 1 |
| 8 | 2.2 | 146.1 | -164.7 | 77.3 | -67.8 | 143.8 | 1.3 | 1 |
| 9 | 2.3 | 137.9 | -162.6 | 169.3 | 70.9 | 114.0 | 1.1 | 0 |
| 10 | 2.4 | 139.1 | -162.5 | 175.6 | -142.3 | 77.8 | 1.0 | 0 |
| 11 | 2.6 | 135.4 | -162.0 | 174.0 | -140.5 | 154.4 | 0.7 | 1 |
| 12 | 2.7 | 144.6 | -160.3 | 74.6 | -69.9 | 99.0 | 0.6 | 1 |
| 13 | 2.7 | 107.1 | -157.9 | 163.1 | -123.2 | 133.8 | 0.6 | 1 |
| 14 | 3.0 | 135.5 | -161.7 | 174.0 | 88.7 | 84.5 | 0.4 | 0 |

Table 1. Output for the conformational search of **1a** compound. Only conformers within 3 kcal/mol are shown. In listing the number of intramolecular H-bonds, the bond between O6S and O4S hasn't been considered. Energy difference with respect to the global minimum ($\Delta$E) is expressed in kcal/mol.

Figure 17 describes the typologies of intramolecular H-bonds (shown in blue) featured in this molecule. γ-turns (as seen in conformer 4) are rarely seen and extended conformation of the peptide is greatly preferred. This was somewhat unexpected, since there is a known tendency of the force field to overestimate folded conformations such as γ-turns for small peptides.[23]



Conf 1 (global minimum)
ΔE = 0.0 kcal/mol

Conf 2
ΔE = 0.5 kcal/mol

Conf 4
ΔE = 1.7 kcal/mol

Figure 17. Representative low-energy conformations (within 2 kcal /mol from the global minimum) from the conformational search of **1a**

In the graph below (Figure 18) it can be seen that β-strands (-135, 135) are much more common than γ-turns (70, 60). Also, no α-helix (-60,-45) and PPII (-75,150) arrangements, usually very common in small peptides, are detected.

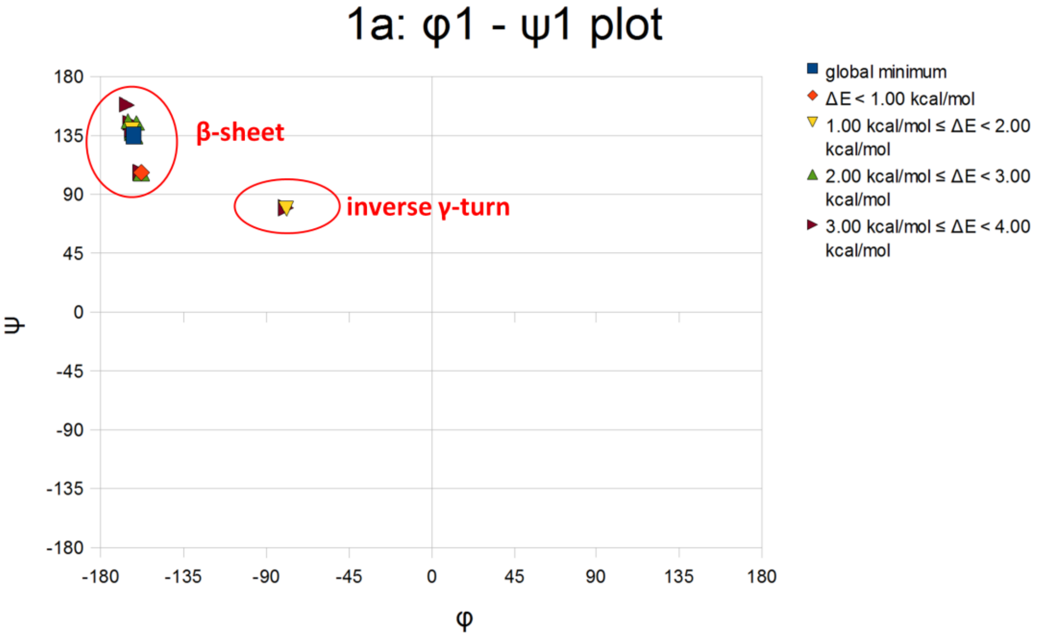


Figure 18. Ramachandran plot for the φ-ψ Asn dihedrals in compound **1a**

Also, no correlation between $\varphi_s$ values and the typologies of intramolecular H-bond featured in the conformer could be found.

The Asn side chain shows a strong preference for the *anti* – *anti* conformation (Figure 19). In the first 2 kcal/mol only conformer no. 5 has a $\chi^2$ angle rather different (*gauche*(+) conformation). The only *gauche*(+) – *gauche*(+) conformation has been found at very high energy, relatively to the global minimum. This is to be expected, since they are generally forbidden for torsional reasons.[24]

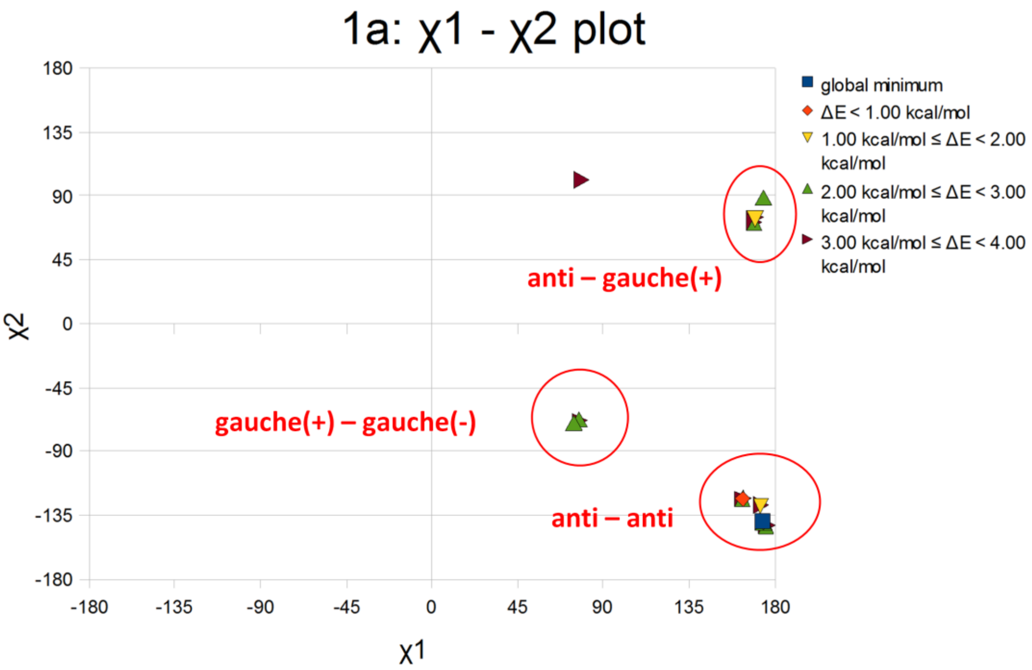
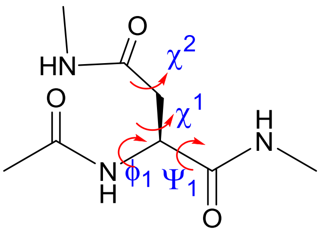


Figure 19. Ramachandran-like plot for the Asn side chain dihedrals in compound **1a**.

It's interesting to see that the $\chi^1$ dihedral is, in the first 2 kcal/mol, always in the *anti* conformation. To see if this is a peculiar feature of this compound, obtained by means of H-bonds between the Asn residue and the sugar moiety, the same MC/EM approach was used to test molecule **1b** (Figure 20) where the Gal has been substituted with a methyl group, thus preventing the formation of any hydrogen bonds comprising the sugar moiety.

### *4.2.1.1 Conformational search: 1b*



**1b**

Figure 20. Compound **1b**. Dihedral angles are defined as in molecule **1a**.

The same options described for the **1a** MC/EM calculation were used. A total number of 7 unique conformers were found in 5 kcal/mol from the global minimum, of which 1 was in the first kcal/mol and 2 in the first 2 kcal/mol.

| Conf | ΔE (kcal/mol) | $\chi^1$ | $\chi^2$ | $\phi_1$ | $\psi_1$ |
|---|---|---|---|---|---|
| 1 | 0.00 | 175.58 | -140.92 | -162.69 | 139.76 |
| 2 | 1.20 | 170.58 | 76.55 | -162.60 | 138.45 |
| 3 | 2.20 | 174.13 | -154.62 | -77.75 | 108.94 |
| 4 | 3.63 | 174.04 | 84.63 | -71.87 | 126.37 |
| 5 | 3.75 | 77.55 | 108.84 | -165.52 | 157.77 |
| 6 | 3.95 | 76.61 | -89.01 | -164.27 | 153.01 |
| 7 | 4.83 | 172.37 | -145.67 | 179.66 | -89.91 |

Table 2. Output for the conformational search of **1a** compound. Only conformers within 5 kcal/mol are shown.

The preference for the extended conformation is confirmed for this peptide, as the global minimum has a much more favorable energy than other conformations. At 2.20 kcal/mol from the global minimum a conformer with a set of φ-ψ angles somewhat in the middle between a PPII and an inverse γ-turn arrangement was found. No proper γ-turn conformation could be obtained.

The $\chi^1$-$\chi^2$ plot for **1b** (see Appendix A.1) doesn't really show any particular difference from the **1a** plot. The *anti* – *anti* is the preferred conformation, and we also saw an *anti* – *gauche*(+) conformation at favorable energy.

In order to confirm this finding we looked at the available rotamer libraries, constructed on the basis of the existing crystallographic structures of polypeptide chains.

The Dunbrack rotamer library[24] shows no strong preference for the *anti* conformation of χ1: the frequency for the *anti* conformation (180°) is pretty much the same as the one for *gauche*(-) conformation. *Gauche*(+) conformation is less favored (Figure 21).

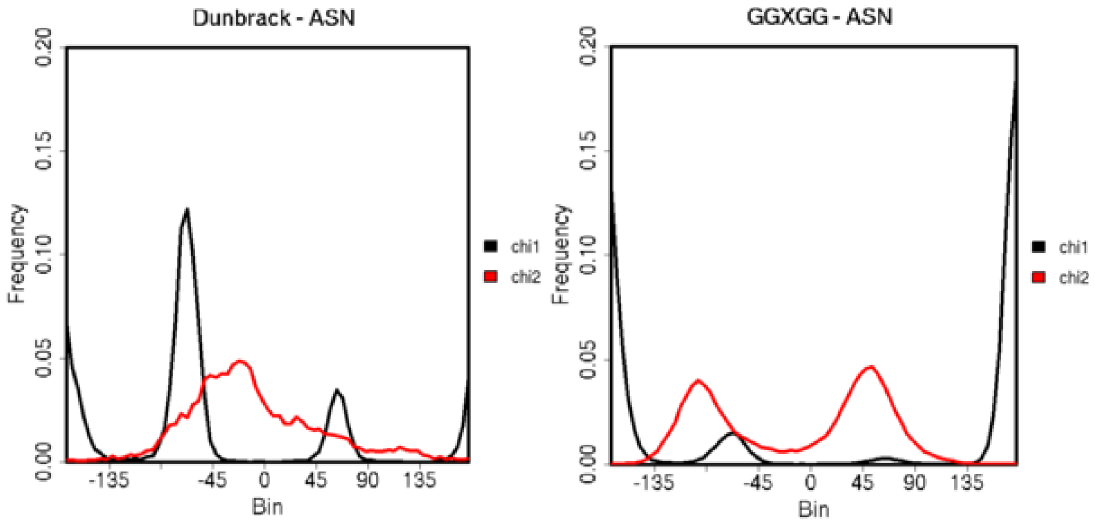
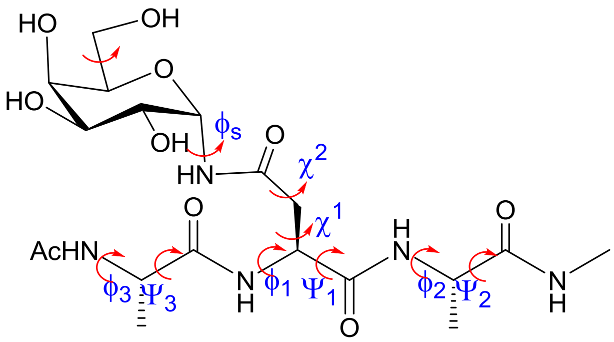


Figure 21. (Left) Dunbrack library for the preferred $\chi^1$-$\chi^2$ values for Asn side chain. (Right) Dynamic simulation of GGNGG (100 ns) in TIP3P water.[25]

On the contrary, the computational study performed by Daggett group,[25] in which a small peptide is simulated for 100 ns in explicit water, shows most of the time an *anti* conformation for $\chi^1$ angle, a characteristic we also found in our conformational searches. However, a different result was obtained for the $\chi^2$ angle. Crystal data are spread over the entire conformational space, with only a slightly more favored *gauche*(-) conformation. Computational data for $\chi^2$ are equally divided between *gauche*(+) and *gauche*(-) conformations. On the contrary, our data are consistent with *anti* conformations at the lowest energies, and only at less favorable energies we see *gauche*(+) and *gauche*(-) arrangements.

### 4.2.2 Conformational search: 2a

A total number of 10 rotatable bonds were varied during the MC/EM calculation.

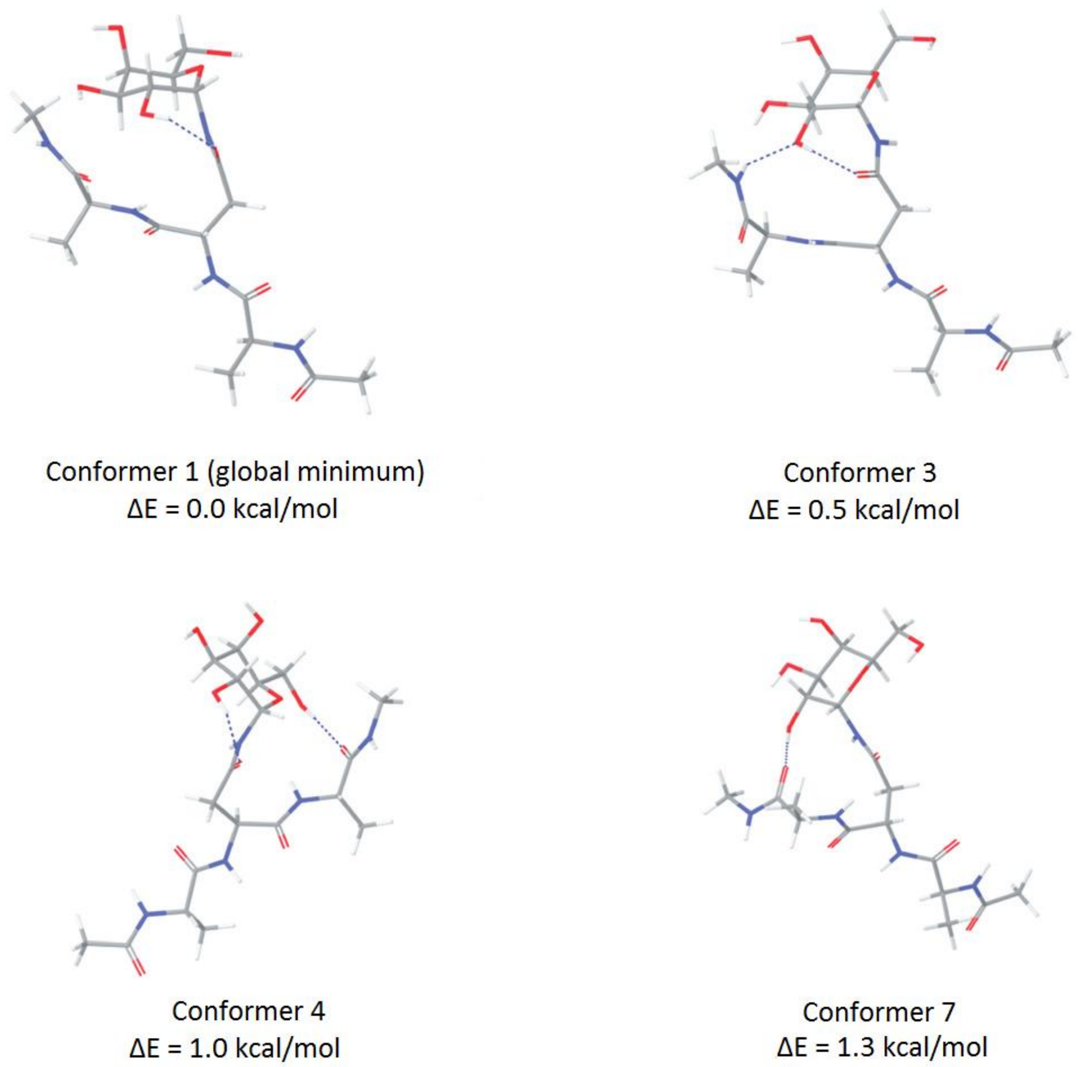


**2a**

Figure 22. Rotatable bonds varied during the conformational search for compound **2a**.

Dihedral angles are defined following the same settings of **1a** compound. The same options described for **1a** computation were used. After multiminimization, a total number of 167 unique conformations were found in 5.00 kcal/mol from the global minimum, of which 4 in the first kcal/mol, 13 in 2.00 kcal/mol and 36 in in 3.00 kcal/mol.

The pictures for conformers 1 (global minimum), 3,4 and 7 are shown in Figure 23 to illustrate the different kinds of H-bonds seen in the most favorable conformers (first 2 kcal/mol).



Conformer 1 (global minimum)
ΔE = 0.0 kcal/mol

Conformer 3
ΔE = 0.5 kcal/mol

Conformer 4
ΔE = 1.0 kcal/mol

Conformer 7
ΔE = 1.3 kcal/mol

Figure 23. Representative low-energy conformations (within 2 kcal/mol) from the global minimum) from the conformational search of **2a**, illustrating the extended conformation of the peptide and the H-bond interactions predicted to occur between the sugar and the peptide chain.

From the analysis of the various φ-ψ couples of **2a** it can be easily verified that almost all the conformers are in extended conformation. This feature, which is in agreement with the NMR data (see section 4.4), could be favored by the presence of H-bonds between the sugar and the Ala residues, forcing them not to assume a more folded conformation. Figure 24 shows the φ-ψ values for the central Asn residue. The distribution of dihedral values for the Ala residues is similar.

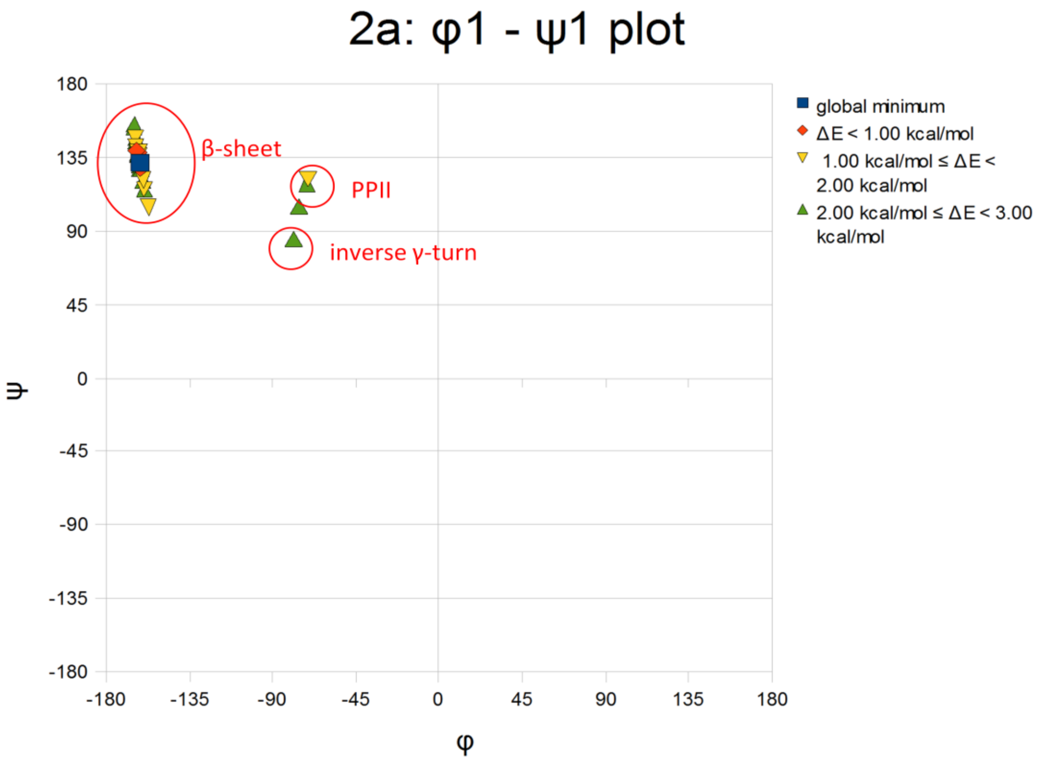
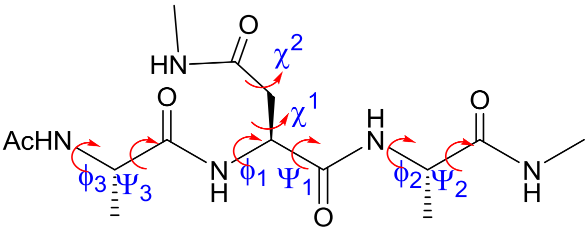


Figure 24. Ramachandran plot for the $\varphi_1$-$\psi_1$ Asn dihedrals in compound **1a**.

The $\chi^1$-$\chi^2$ plot shows a limited range of conformations for **2a**. The *anti* – *anti* conformation is energetically favored, and thus more populated. The only other conformation found is the *anti* – *gauche*(+), though now at much more favored energies (see Appendix A.2).

### *4.2.2.1 Conformational search: 2b*



**2b**

Figure 25. Compound **2b**. Dihedral angles are defined as in molecule **1a**.

The conformational search for **2b** (a slightly modified Ala-Asn-Ala tripeptide) was done mostly to see if the extended conformation adopted by **1a** and **2a** was due to the presence of the sugar or if it is a natural propensity of Asn-derived small peptides.

| Conf | ΔE | $\chi^1$ | $\chi^2$ | $\phi_1$ | $\psi_1$ | $\phi_2$ | $\psi_2$ | $\phi_3$ | $\psi_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 175.43 | -136.12 | -163.95 | 146.01 | -159.49 | 138.65 | -162.32 | 145.58 |
| 2 | 0.99 | 169.00 | 69.64 | -163.88 | 143.83 | -159.65 | 142.39 | -162.34 | 145.58 |
| 3 | 1.60 | 172.36 | -151.75 | -161.68 | 132.37 | -144.72 | 34.97 | -162.34 | 145.60 |
| 4 | 1.69 | 172.75 | 134.20 | -163.17 | 140.33 | -78.86 | 73.80 | -162.33 | 145.59 |
| 5 | 2.12 | 173.98 | -155.05 | -74.25 | 111.32 | -162.59 | 145.35 | -162.42 | 146.11 |
| 6 | 2.16 | 175.36 | -136.06 | -164.54 | 146.44 | -159..39 | 138.48 | -71.14 | 123.00 |
| 7 | 2.17 | 173.94 | -149.01 | -71.52 | 118.96 | -161.21 | 143.01 | -162.40 | 145.94 |
| 8 | 2.19 | 174.58 | -96.63 | -162.92 | 141.62 | -143.01 | 57.62 | -162.34 | 145.60 |
| 9 | 2.23 | 178.20 | -146.86 | -163.81 | 146.50 | -70.50 | 121.99 | -162.32 | 145.57 |

Table 3. Output for the conformational search of **2b**. Energy difference with respect to the global minimum (ΔE) is expressed in kcal/mol. Only the conformers within 3.00 kcal/mol are shown.

From the analysis of the results obtained from the calculation, summarized in Table 3, we couldn't see any relevant change in the backbone dihedral values distribution. In the first 2.00 kcal/mol from the global minimum, φ–ψ values for the Ala-Asn-Ala residues are again in β-strand conformation.

Is then the β-strand the natural conformation for the Ala-Asn-Ala tripeptide, meaning that the sugar is not responsible for it? Do all tripeptides AXA remain in a β-strand? No literature data exist to answer the first question. Despite that, other AXA tripeptides have been studied and are known to be mostly in extended conformation. A study published in 2004[26] showed that AXA tripeptides (X being valine, tryptophan, histidine, and serine) predominantly adopt an extended β-strand conformation while AXA tripeptides for which X is lysine and proline prefer a polyproline II-like (PPII) structure. In turn, other studies[27] demonstrated that the X residue in the AXA peptide is sometimes responsible for the folding of the molecule. Our findings seem to indicate that glycosylation has little to none effect on the conformation of the peptide backbone conformation.

### 4.2.3 Mixed Monte Carlo Metropolis / Stochastic Dynamics

**1a**, **1b**, **2a** and **2b** were also subjected to MC/SD methods to better explore the conformational space of such molecules.

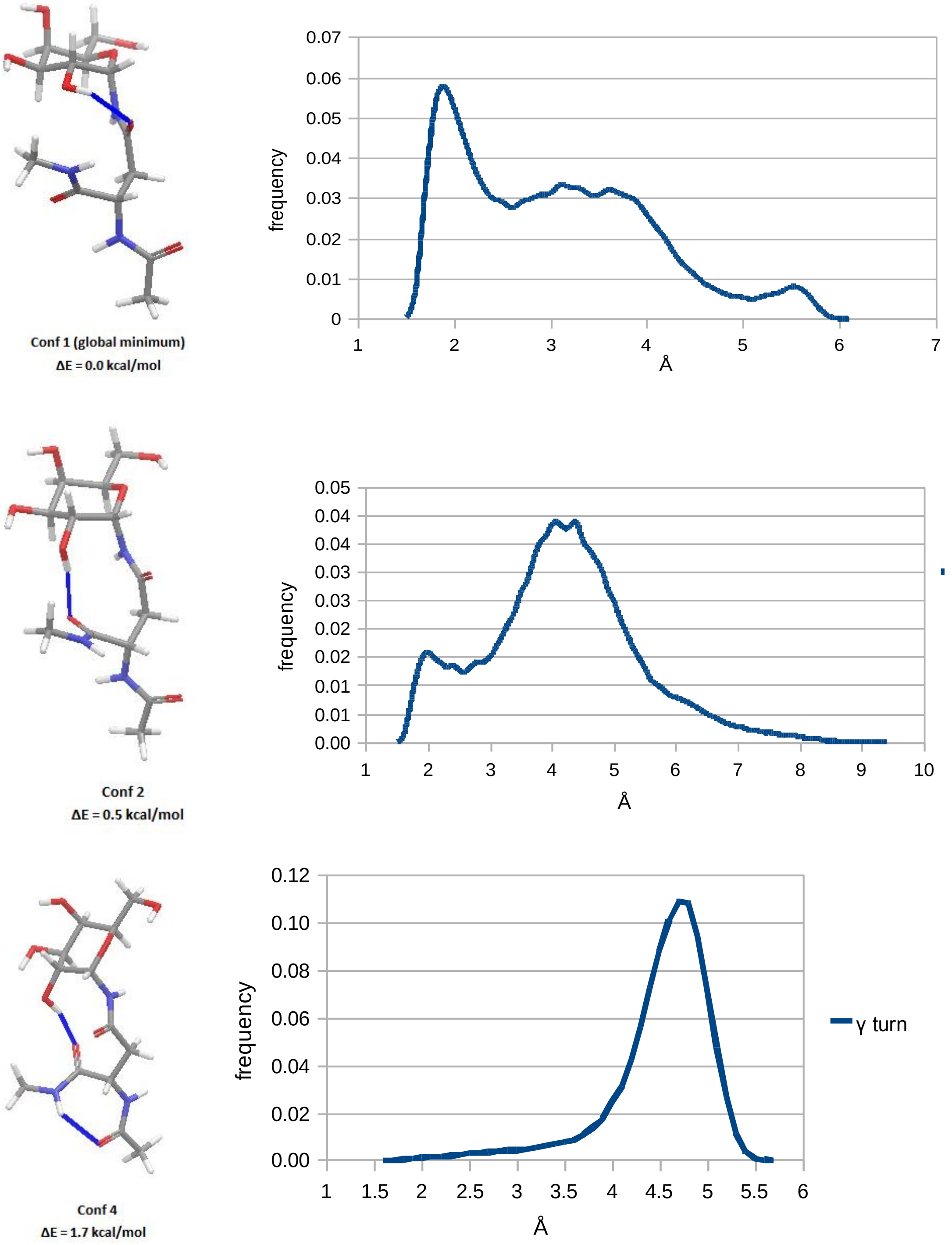
The simulation time was 5 ns for **1a** and **1b** compounds, and 10 ns for **2a** and **2b** compounds. MacroModel, from Schrodinger, was used to perform these calculations.

#### *4.2.3.1 MC/SD: 1a*

The same torsions described in Figure 16 were varied during the run. The MC acceptance ratio was about 4.5 %. Both the torsions and the hydrogen bond distances, as seen in the different conformers of the conformational search, were monitored.

In Figure 26 the distribution of the distances between the atoms forming the relevant hydrogen bonds are plotted. The distance between the Gal O2 and the carbonyl group of the Asn side chain has a higher probability of being at ca. 2 Å, with higher distance values having subsequent less frequency. The distance distribution between the Gal O2 and the carbonyl group of the C-terminal moiety has two peaks, only one allowing the formation of the H-bond (which is depicted in Figure 26, conformer 2). The H-bond forming the 7-atom ring between H-N- and O=C in the Asn residue was seen only in the 0.66 % of the time. This is also confirmed by the distribution of the related distances, as the peak of the curve is around 4.7 Å. All together, these findings confirm what was already pointed out by the conformational search of **1a**. Moreover, the distribution of the dihedral angles $\varphi$, $\psi$, $\chi^1$ and $\chi^2$ over time corresponds to those found in the previous study (see Appendix A.3). $\varphi$ and $\psi$ angles are such that β-sheet conformation is predominant. A small percentage has a $\varphi$ angle around -80 which, coupled with a $\psi$ angle of ca. 70, brings an inverse γ-turn.

MC/SD data for **1b**, not shown here, found, as previously, a prevalence of extended conformations.



Conf 1 (global minimum)
ΔE = 0.0 kcal/mol

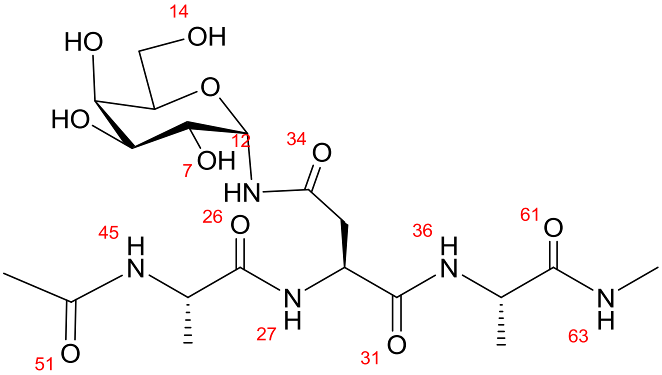
Conf 2
ΔE = 0.5 kcal/mol

Conf 4
ΔE = 1.7 kcal/mol

Figure 26. Distribution of the distance (right) between the atoms forming the hydrogen bond highlighted in blue in the 3D structure (left).

### *4.2.3.2 MC/SD: 2a*

Nine torsions were considered for the dynamics simulation of **2a**. The MC acceptance ratio was about 4.0 %. 9 possible H-bonds were monitored, along with the $\chi^1$, $\chi^2$ and $\varphi - \psi$ dihedrals. In the table below, the atom pairs of which we monitored the distance are shown, together with a percentage of structures with the proper H-bond (see the 2D structure for the atom labels).



| Atom pair | hydrogen bond | % of occurence |
|---|---|---|
| 12 – 31 | O2 – C=O (backbone Asn) | 8.98 |
| 12 – 34 | O2 – C=O (side chain Asn) | 25.63 |
| 12 – 61 | O2 – C=O (C-term) | 2.12 |
| 14 – 61 | O6 – C=O (C-term) | 1.61 |
| 27 – 51 | partial γ-turn (N-term) | 0.45 |
| 26 – 36 | γ-turn | 0.63 |
| 45 – 61 | β-hairpin | 0.00 |
| 7 – 63 | O2 – NH (C-term) | 8.67 |
| 31 – 63 | partial γ-turn (C-term) | 0.19 |

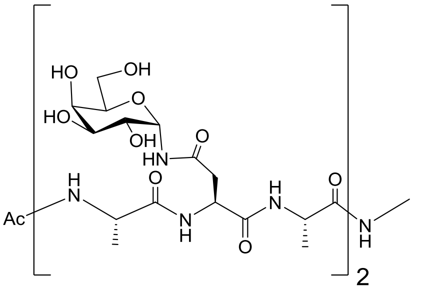
Table 4. Percentage of occurrence of an hydrogen bond between a list of atom pairs in the MC/SD run of **2a**.

From Table 4 it is quite clear that this substituted tripeptide mostly adopts an open conformation: turns almost never appear during the MC/SD run. As with **1a**, an H-bond between O2-Gal and the carbonyl group of the Asn side chain (atom pair 12-34) is abundantly present. An extended conformation was also found for **2b**.

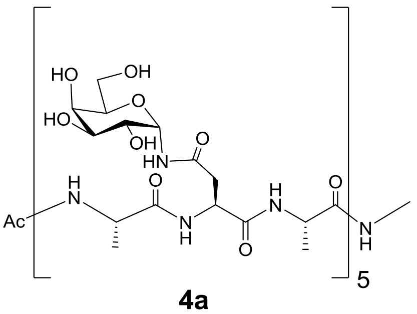
In conclusion we found that, for both **1a** and **2a**, the extended conformation is greatly preferred and intramolecular H-bonds between hydroxyl groups of the galactose moiety and acceptor groups in the Asn side chain moiety are present in the most energetically favored conformations. These results were confirmed by NMR experiments performed by F. Marcelo from the group of prof. Jiménez-Barbero (Consejo Superior de Investigaciones Científicas, CSIC) in Madrid (see section 4.4).[18]

## 4.3 CONFORMATIONAL ANALYSES OF 3a AND 4a

Molecules **3a** and **4a** possess a high number of degrees of freedom which makes impossible for them to be analyzed by a full conformational search.



**3a**



**4a**

Rather, we devised the following protocol:

- a short (10 ns) MD simulation with an explicit TIP3P water solvent, using the AMBER 9 package,[28] followed by
- four simulated annealing (SA) simulations (where molecules are rapidly heated and then cooled in a controlled way) performed in an implicit model solvation, using MacroModel.

The final structure of the MD simulation was the starting structure for the first SA. The same goes for the SA runs, where the final conformation of the preceding run was the starting point for the subsequent one.

The first step in the protocol was done in an explicit solvent mostly to test whether changing the solvation model would have had any major impact on the conformation of the starting structure.

### 4.3.1 Results: 3a

An initial simulation (using AMBER 9) of 10 ns was performed for **3a**, starting from an extended conformation. In Figure 27 the Root Mean Square Deviation of each saved frame from the MD trajectory is shown.

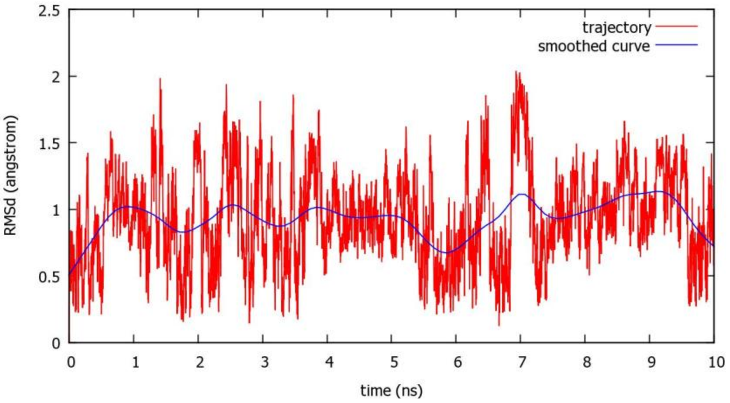


Figure 27. RMSD from the starting structure (in extended conformation) for **3a** over 10 ns of MD.

The RMSD is computed considering only the heavy atoms of the backbone. For the whole simulation the RMSD never reaches a value greater than 2 Å, which is a good indication of the fact that no conformational change is occurring. Even in this case, with an explicit solvent, an extended conformation was maintained during the simulation.

The final structure of the MD simulation was then stripped of the water molecules and used as the starting point for the simulated annealing. Four SA simulations have been performed and the final structure obtained from the fourth simulation is shown in Figure 28. The final conformation has kept the β-strand arrangement, thus confirming the results obtained from the simulations performed in explicit water.

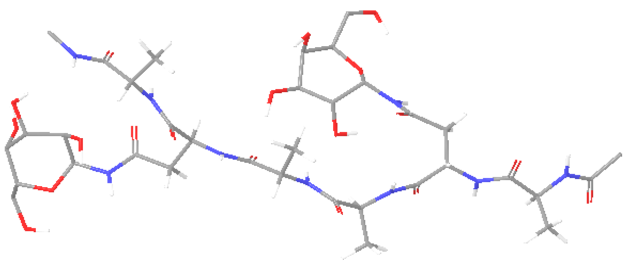


Figure 28. Final structure obtained from the simulated annealing protocol used for **3a**.

Due to the increased size of **3a**, monitoring each significant hydrogen bond and dihedral angle does not give a complete picture of the dynamical behavior of **3a** during the SA simulations. As a consequence we chose to monitor all the φ-ψ values sampled by each residue during the four SAs and then construct a 3D Ramachandran plot. To do it, the 2D φ-ψ map is divided in an equally spaced grid. Afterwards, each sampled φ-ψ value is assigned to a point in the grid and counted. The Python script which has been written to implement the algorithm is contained in Appendix A.4. The result is a 3D visualization of the frequency with which regions in the φ-ψ are sampled, and is shown in Figure 29.

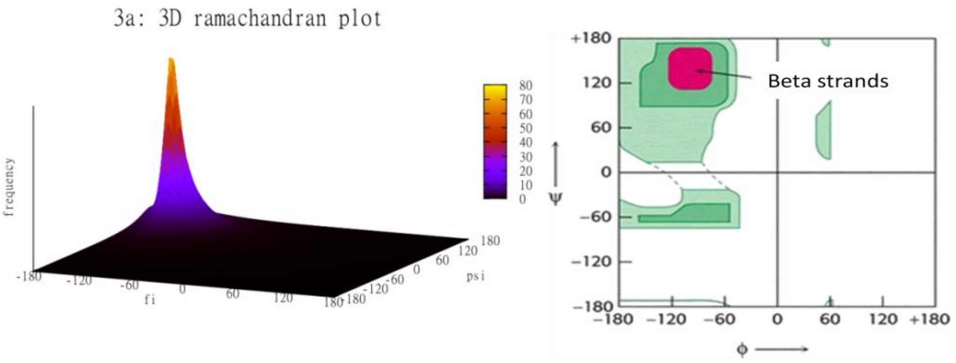


Figure 29. 3D and 2D Ramachandran plot for **3a**.

The Ramachandran plot shows that, for the great majority of the time, even at the high temperature required by the simulated annealing only the β-strand region was sampled. This was again a confirmation of the already pointed out conformational stability of **3a**.

The predicted β-strand arrangement for **3a** was confirmed by NMR experiments performed by Dr. F. Vasile from the University of Milan (see section 4.4).

### 4.3.2 Results: 4a

Due to the increased size of **4a** (n=5) with respect to **2a** (n=2) the initial MD simulation in explicit water was elongated to 50 ns. Starting from an extended conformation, we observed a sudden conformational change, leading to a more globular arrangement, suggesting that in this case a β-strand backbone conformation was no longer the preferred one, at least in an explicit water environment. In Figure 30 the RMSD (calculated using the backbone heavy atoms) from the initial structure is shown, for the 50 ns MD run. The conformational transition can be observed to occur during the first 5 ns of simulation. A second, smaller, transition, occurring at ca. 35 ns, involves a loop movement near the N-terminus, resulting in the N and C-terminus being more close.

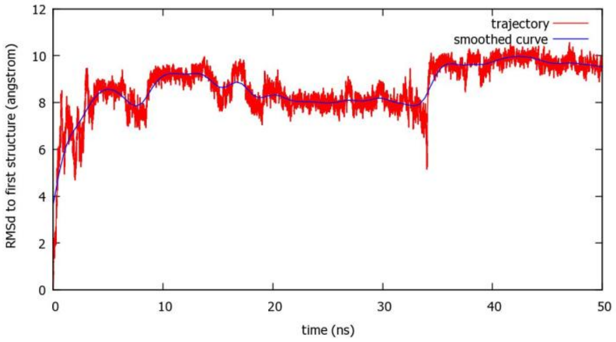


Figure 30. RMSD from the starting structure (in extended conformation) for **4a** over 50 ns of MD.

The radius of gyration computed for each frame during the simulation (Figure 31) shows a similar behavior. The drop in the radius of gyration value over the first 5 ns, shows that the conformational change is directed towards a more folded conformation, and so does the second, smaller drop at 35 ns. However, no stable folded conformation could be found. This suggested that a random coil conformation was adopted during this calculation. In fact, the lowest energy structure extracted from the MD simulation, does not show any clear secondary structure features. (Appendix A.5).

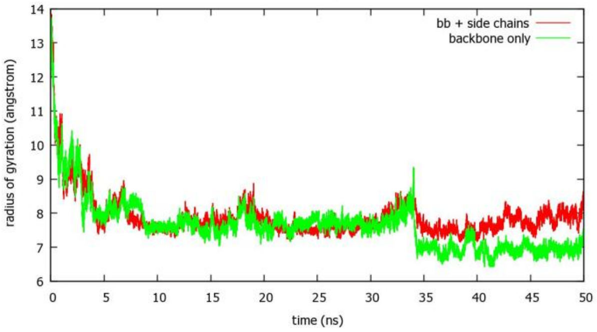


Figure 31. RMSD (using the backbone heavy atoms, in red, and the backbone and side chain heavy atoms, in green) from the starting structure (in extended conformation) for 4a over 50 ns of MD.

Using the same protocol developed for **3a**, the final structure of the MD simulation was used to start the first run of the 4 SA experiments. In Figure 32 the final conformer obtained from the SAs is shown.

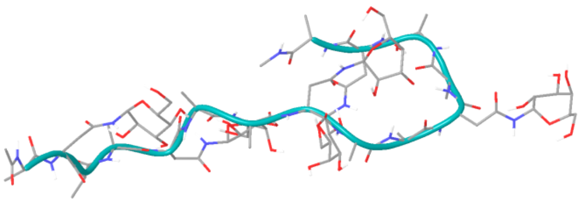


Figure 32. Final structure obtained from the fourth simulated annealing run for **4a**.

The structure is not fully folded: partial α-helix and turns can be observed, but no clear secondary arrangement is present. The analysis of the backbone dihedral angles, using the same protocol developed for **3a**, led to a distribution of values as shown in the 3D Ramachandran plot of Figure 33.

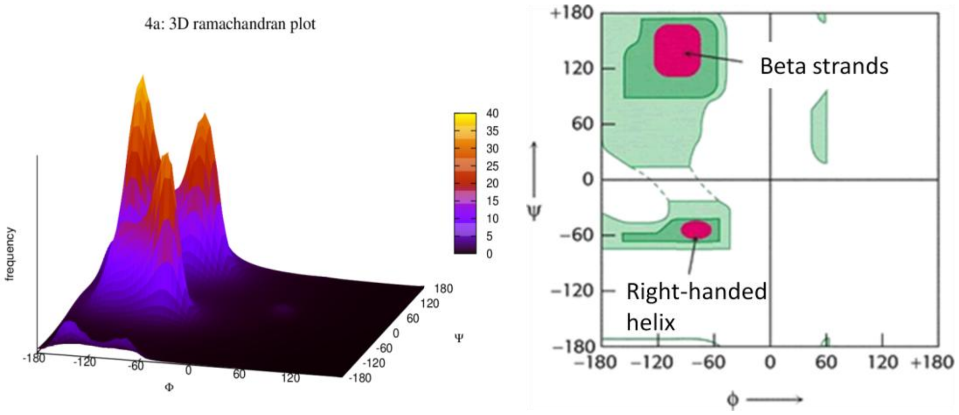


Figure 33. 3D and 2D Ramachandran plot for **4a**.

The 3D Ramachandran plot shows that in this case, β-strands and α-helices dihedral values are equally present during the simulations, indicating that no clear secondary structure is formed and a more disordered, random coil conformation is the most probable one.

## 4.4 Experimental validation of results

The solution conformations of the two α-*N*-linked glycopeptides **1a** and **2a** were investigated by NMR spectroscopy by F. Marcelo, from Prof. Barbero (CIC-CSIC, Madrid) group. Coupling constants and NOE data were determined and their analysis allowed to evaluate the conformation of the peptide backbone in water solution (Figure 34).[18]

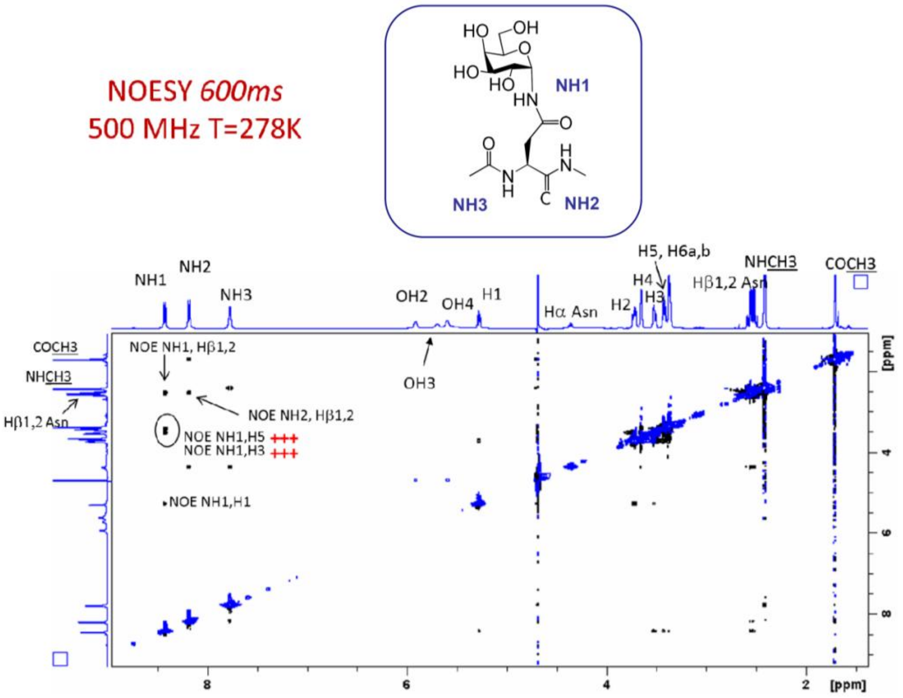


Figure 34. 2D-NOESY spectrum obtained for glycopeptide **1a** in H2O/D2O 90:10 recorded at 500MHz with 600ms of mixing time and at 278K.

The experimental $^3J_{NH,H\alpha}$ coupling constants values strongly indicated the presence of an extended conformation for the peptide backbone in solution, while the J Hα,Hβ1 / J Hα Hβ2 values (5.1/6.5 Hz and 8,1/7.3 Hz, for **1a** and **2a**, respectively) revealed the existence of certain flexibility around 1 (H-C-C-H) of both Asn residues. The absence of non-vicinal medium-range NOE contacts around $\chi^1$ suggest that glycopeptide **2a** adopts, as its main conformation, an extended conformation of the peptide backbone when free in solution (Figure 35).

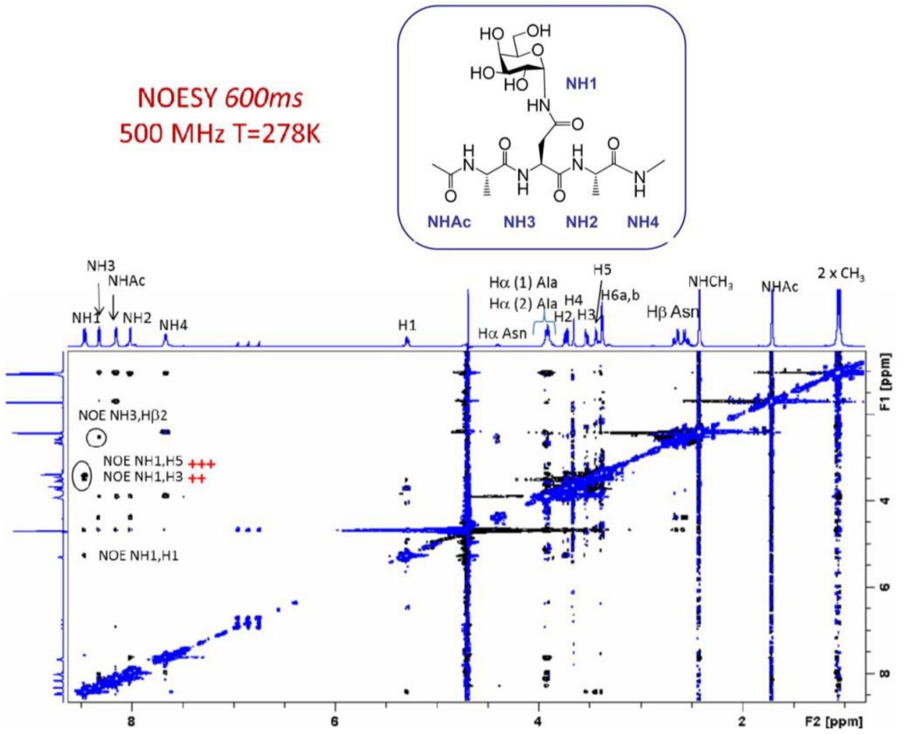
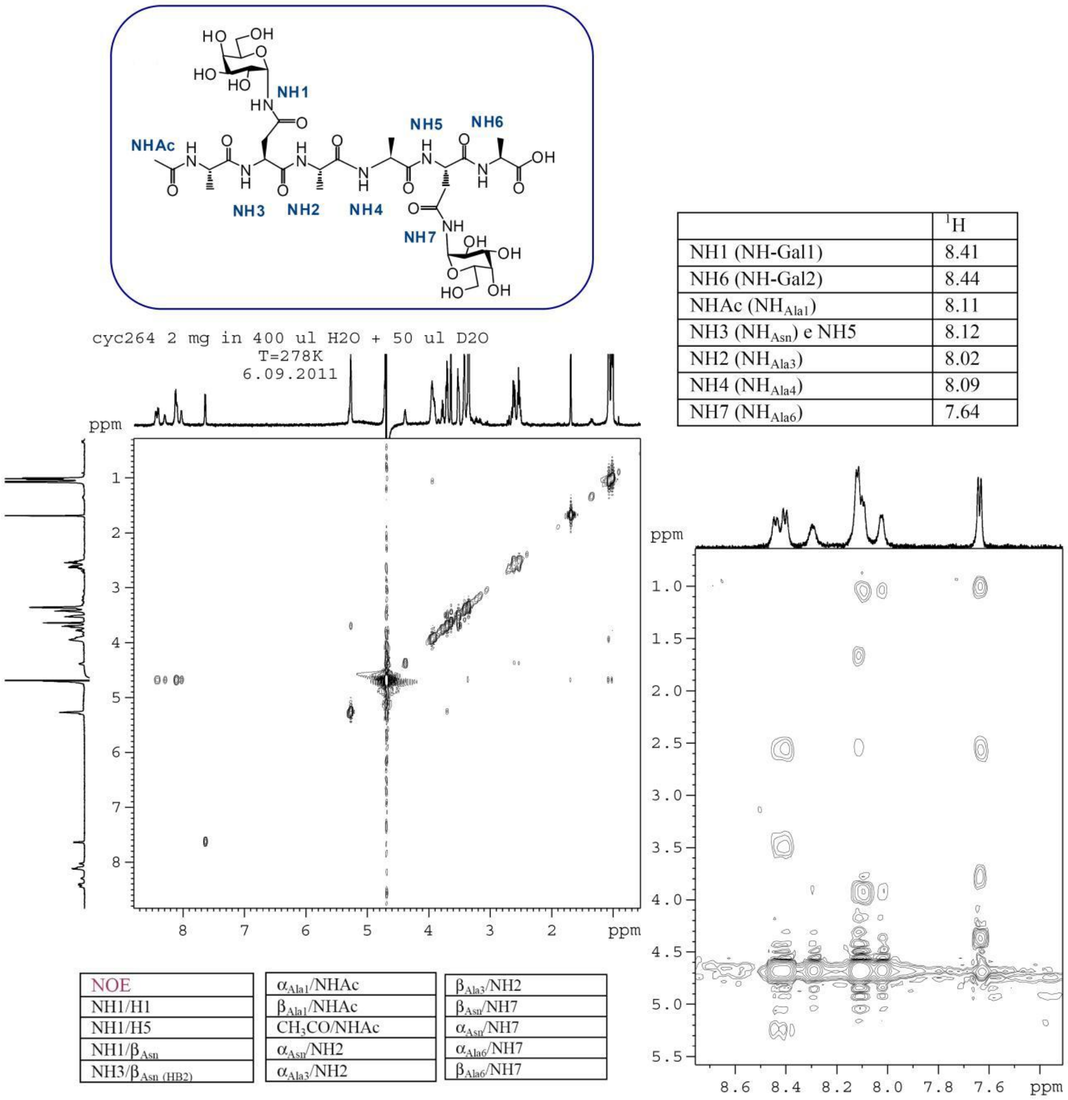


Figure 35. 2D-NOESY spectrum obtained for glycopeptide **2a** in $H_2O/D_2O$ 90:10 recorded at 500MHz with 600ms of mixing time and at 278K.

Molecule **3a** was also studied[d] (Figure 36) and the two subunits were found to have overlapping signals. Also for this molecule the extended conformation suggested by our calculations was confirmed by the $^3J_{NH,H\alpha}$ coupling constants and by the absence of non-vicinal medium range NOE contacts around $\chi^1$.

[d] Dr. F. Vasile, University of Milan, data not published



| | $^1$H |
|---|---|
| NH1 (NH-Gal1) | 8.41 |
| NH6 (NH-Gal2) | 8.44 |
| NHAc (NH$_{Ala1}$) | 8.11 |
| NH3 (NH$_{Asn}$) e NH5 | 8.12 |
| NH2 (NH$_{Ala3}$) | 8.02 |
| NH4 (NH$_{Ala4}$) | 8.09 |
| NH7 (NH$_{Ala6}$) | 7.64 |

| NOE | $\alpha_{Ala1}$/NHAc | $\beta_{Ala3}$/NH2 |
|---|---|---|
| NH1/H1 | $\beta_{Ala1}$/NHAc | $\beta_{Asn}$/NH7 |
| NH1/H5 | $CH_3CO$/NHAc | $\alpha_{Asn}$/NH7 |
| NH1/$\beta_{Asn}$ | $\alpha_{Asn}$/NH2 | $\alpha_{Ala6}$/NH7 |
| NH3/$\beta_{Asn\ (HB2)}$ | $\alpha_{Ala3}$/NH2 | $\beta_{Ala6}$/NH7 |

Figure 36. 2D-NOESY spectrum for glycopeptide **3a** in H2O/D2O recorded at 600MHz at 278K.

For the moment, no NMR data are available to confirm the computational findings regarding glycopeptide **4a**, but the analyses will be performed on it, as soon as additional material will be available.

## 4.5 Conclusions

Computational modelling techniques such as conformational search and molecular dynamics have proven to be an invaluable tool for predicting the 3D properties of small glycopeptides. In this chapter, I have shown that simulations can predict experimental data in a completely independent way, without any experimental constrain applied during the computations.

From the results of **4a** it seems that with increasing size, the extended conformation, conserved in all small glycopeptides analyzed, is lost. No experimental data is available to support this prediction. However, Circular Dichroism (CD) spectra performed on a similar glycopeptide (where the tripeptide Ala-Asn-Ala is repeated four times, instead of five) showed random coil as the most preferred conformation.[29]

In any case, more studies are needed in order to get a full understanding of the conformational space sampled by this rather large compound. Longer simulations will also be performed to try and predict whether these molecules possess antifreeze activity, of which the experimental test is ongoing.

## 4.6 Methods

MC/EM and MC/SD calculations were performed using MacroModel[21] and the Maestro[30] Graphical User Interface. The AMBER* force field with the Senderowitz-Still parameters[31] has been used. Water solvation was simulated using a GB/SA continuum solvent model.[32]

For MC/EM, Extended non-bonded cut off distances (a van der Waals cut off of 8.0 Å and an electrostatic cutoff of 20.0 Å) were used. See Appendix A.6 for the command files used to perform the calculations. 1000 MC steps per rotatable bond were applied. This ensured the convergence of the simulation.

For MC/SD, following prior studies,[33] Van der Waals, electrostatic and H-bond cutoff were extended to 25 Å, 25 Å and 15 Å, respectively. Compounds were equilibrated for 1.0 ps prior the actual MC/SD run. The calculations were performed at 300 K with a timestep of 1.5 fs; no SHAKE algorithm has been used. A total of 5000 snapshots were saved during the run. In every MC step, the number of torsions to be changed randomly varied from 1 to n (n being the total number of bonds considered as rotatable). For every compound, 2 different runs were performed, using different starting points, to verify that all of the conformational space was sampled. In addition, snapshots saved during the run were again minimized, following prior literature,[33] as this protocol proved to be more in agreement with the experimental data.

Molecular Dynamics (MD) simulations were performed using the AMBER 9 package[28] with the ff99sb[34] force field assisted by the glycam04 parameters[35] for the Galactose moiety and the glycosidic linkage. Before starting the production runs, compounds were minimized and allowed to relax in a cubic box with fixed volume, while gently reaching the desired value of 300K. A weak constraint on the solute was applied at this stage. A 200 ps run at constant pressure completed the equilibration step.

Simulated Annealing experiments were done using MacroModel[21] by performing each time 10 ns of MD. in which the initial structure is brought at the temperature of 500 K and then uniformly cooled to 50 K over the entire simulation. A time step of 1.5 ns was used. A continuum solvent model and extended non bonded interactions were utilized (see Appendix A.7 for **3a** and **4a** command files).

## 4.7 BIBLIOGRAPHY

(1) DeVries, A. L.; Vandenheede, J.; Feeney, R. E. *J. Biol. Chem.* **1971**, *246*, 305–308.

(2) Tam, R. Y.; Rowley, C. N.; Petrov, I.; Zhang, T.; Afagh, N. A.; Woo, T. K.; Ben, R. N. *J. Am. Chem. Soc.* **2009**, *131*, 15745–15753.

(3) Garner, J.; Harding, M. M. *Chembiochem* **2010**, *11*, 2489–2498.

(4) Wang, J. H. *Cryobiology* **2000**, *41*, 1–9.

(5) Tachibana, Y.; Fletcher, G. L.; Fujitani, N.; Tsuda, S.; Monde, K.; Nishimura, S.-I. *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 856–862.

(6) Hays, L. M.; Feeney, R. E.; Crowe, L. M.; Crowe, J. H.; Oliver, A. E. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 6835–6840.

(7) Inglis, S. R.; Turner, J. J.; Harding, M. M. *Curr. Protein Pept. Sci.* **2006**, *7*, 509–522.

(8) Carpenter, J. F.; Hansen, T. N. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 8953–8957.

(9) Chao, H.; Davies, P. L.; Carpenter, J. F. *J. Exp. Biol.* **1996**, *199*, 2071–2076.

(10) Bouvet, V.; Ben, R. N. *Cell Biochem. Biophys.* **2003**, *39*, 133–144.

(11) Leclère, M.; Kwok, B. K.; Wu, L. K.; Allan, D. S.; Ben, R. N. *Bioconjug. Chem.* **2011**, *22*, 1804–1810.

(12) Balcerzak, A. K.; Ferreira, S. S.; Trant, J. F.; Ben, R. N. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 1719–1721.

(13) Wilkinson, B. L.; Stone, R. S.; Capicciotti, C. J.; Thaysen-Andersen, M.; Matthews, J. M.; Packer, N. H.; Ben, R. N.; Payne, R. J. *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 3606–3610.

(14) Colombo, C. Ph.D Thesis, Synthesis of unnatural α-N-linked glycopeptides, University of Milan, 2011.

(15) Bosques, C. J.; Tschampel, S. M.; Woods, R. J.; Imperiali, B. *J. Am. Chem. Soc.* **2004**, *126*, 8421–8425.

(16) Gabius, H. J.; Walzel, H.; Joshi, S. S.; Kruip, J.; Kojima, S.; Gerke, V.; Kratzin, H.; Gabius, S. *Anticancer Res.* *12*, 669–675.

(17) Shaanan, B.; Lis, H.; Sharon, N. *Science* **1991**, *254*, 862–866.

(18) Marcelo, F.; Cañada, F. J.; André, S.; Colombo, C.; Doro, F.; Gabius, H.-J.; Bernardi, A.; Jiménez-Barbero, J. *Org. Biomol. Chem.* **2012**, *10*, 5916–5923.

(19) Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

(20) Guarnieri, F.; Still, W. C. *J. Comput. Chem.* **1994**, *15*, 1302–1310.

(21) MacroModel, version 9.5, Schrödinger, LLC, New York, NY, **2007**.

(22) Corzana, F.; Busto, J. H.; Jiménez-Osés, G.; García de Luis, M.; Asensio, J. L.; Jiménez-Barbero, J.; Peregrina, J. M.; Avenoza, A. *J. Am. Chem. Soc.* **2007**, *129*, 9458–9467.

(23) Gnanakaran, S.; Garcia, A. E. *J. Phys. Chem. B* **2003**, *107*, 12555–12557.

(24) Dunbrack, R. L.; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661–1681.

(25) Van der Kamp, M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R. D.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S.; Bromley, D.; Beck, D. A. C.; Daggett, V. *Structure* **2010**, *18*, 423–435.

(26) Eker, F.; Griebenow, K.; Cao, X.; Nafie, L. A.; Schweitzer-Stenner, R. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 10054–10059.

(27) Motta, A.; Reches, M.; Pappalardo, L.; Andreotti, G.; Gazit, E. *Biochemistry* **2005**, *44*, 14170–14178.

(28) Case, D.A., Darde, T.A., Cheatham III, T.E., Simmerling, C.L., Wang, J., Duke, R.E., Luo, R., Merz, K.M., Pearlman, D.A., Crowley, M., Walker, R.C., Zhang, W., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Wong, K.F., Paesani, F., Wu, X., Brozell, S., Ts, M.; D.H., Schafmeister, C., Ross, W.S., Kollman, P. A. *Univ. California, San Fr.* **2006**.

(29) Stucchi, M. Master Thesis, University of Milan, 2012.

(30) Maestro, version 8.0, Schrödinger, LLC, New York, NY **2007**.

(31) McDonald, D. Q.; Still, W. C. *Tetrahedron Lett.* **1992**, *33*, 7743–7746.

(32) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(33) Brocca, P.; Bernardi, A.; Raimondi, L.; Sonnino, S. *Glycoconj. J.* **2000**, *17*, 283–299.

(34) Hornak, V.; Abel, R.; Okur, A. *Proteins: Struct., Funct., Bioinf.* **2006**, *725*, 712–725.

(35) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10541–10545.

# TARGETING PROTEIN-PROTEIN INTERACTIONS: TYPE I CLASSICAL CADHERINS

# Chapter 5: Cadherins

## 5.1 INTRODUCTION

Modern molecular and cellular biology have enabled enormous progress in unveiling many aspects of the complex network of physiological and pathological mechanism, leading to the discovery of new targets for human therapeutics. In many physiological processes, the role played by protein-protein interactions (PPI) is central, from cell-cell messaging to cellular apoptosis. Targeting the interfaces between proteins has huge therapeutic potential, but developing drug-like small molecules that modulate PPIs is a great challenge.[1] However, impressive progress has been made in the discovery of small organic modulators of PPIs, highlighting how the combined research efforts in the areas of computational modeling, organic synthesis, structural chemistry and biological screening can lead to major advances in the field.[2]

This part of my PhD thesis is based on a wide project[e] involving different research groups (Figure 37) and aimed at applying a multidisciplinary approach to tackle the problem of finding small peptidomimetic inhibitors of cadherins PPIs, a class of adhesive proteins. The project benefits from collaborations with researchers from the Consiglio Nazionale delle Ricerche and the University of Insubria, for the synthesis of the small molecules being designed, with scientists of the Dept. of Experimental Oncology and Molecular Medicine at the Istituto Nazionale Tumori, for the *in vitro* tests of the synthesized molecules, with Dr. Parisini's group at the Istituto Italiano di Tecnologia, for the co-crystallization of cadherin constructs and inhibitors, and with Dr. Potenza's group at the University of Milan for solution NMR studies of the protein-ligand complexes.

[e] Computer-aided design, synthesis and biological evaluation of peptidomimetics targeting N-cadherin as anticancer agents, MIUR-FIRB ‘Futuro in Ricerca’ RBFR088ITV

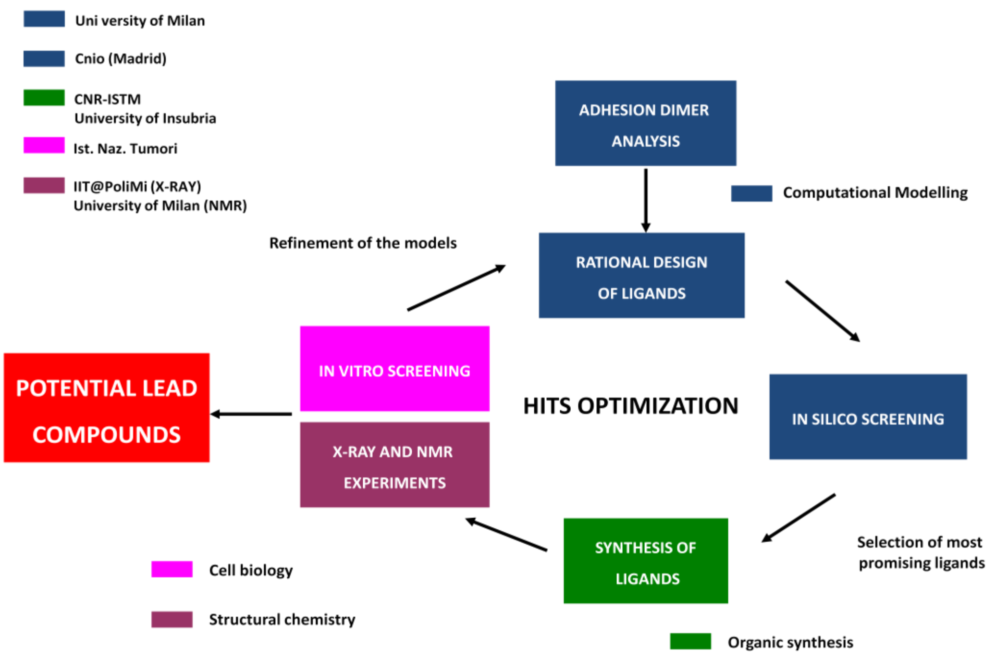


Figure 37. Workflow depicting the various collaborations through which cadherins inhibitors are being designed and tested.

My contribution to the project concerned the application of computational modelling techniques aimed at obtaining:

- a detailed characterization of the E-cadherin and N-cadherin (two members of the so-called classical cadherins) homophilic interaction from available crystal structures, discussed in Chapter 6.
- the set up and validation of in silico screening protocols and the design of small peptidomimetic E- and N-cadherin modulators, discussed in Chapter 7.
- an understanding of the dimerization mechanism of classical cadherins, discussed in Chapter 8 and based on the studies performed during my stay at the Centro Nacional de Investigaciones Oncológicas (CNIO), in Madrid, under the supervision of Prof. Francesco Luigi Gervasio.

In this chapter an overview on classical cadherins  and on their role in the development of cancer will be shortly presented. A more detailed discussion can be found in recently published reviews.[3–7]

## 5.2 CADHERIN SUPERFAMILY AND CLASSICAL CADHERINS

Animal tissues are bound together thanks to adhesive forces in their component cells. Two principal types of adhesive junctions exist: desmosomes and adherens junctions, their role being maintaining cell shape and tissue integrity. Adherens junctions are cell-cell adhesion complexes found in a variety of cells,[8] characterized by opposed plasma membranes with an in-between space of 15 to 30 nm. Desmosomes reinforce adhesion and are mostly present in organs like heart and skin.[9]

Among the adherens junction components, cadherins are the core elements.[10] The level of cadherin expression influences the strength of adhesion, whereas the type of cadherin expressed determines the specificity and the properties of cell interactions. Cadherins found in adherens junctions were the first members of a broader superfamily of cadherins to be discovered, and thus are now being called classical cadherins. Cadherins can be classified into subfamilies based on the number and arrangement of their Extra Cellular (EC) domains (Figure 38), common structural components of Ig-like fold comprising ca 110 amino acids and numbered according to their distance from the membrane, being EC1 the N-terminal membrane-distal domain. Most EC domains contain conserved $Ca^{2+}$ ions[11] in the linker regions, in order to rigidify the EC structure[12] and avoid proteolysis.[10]

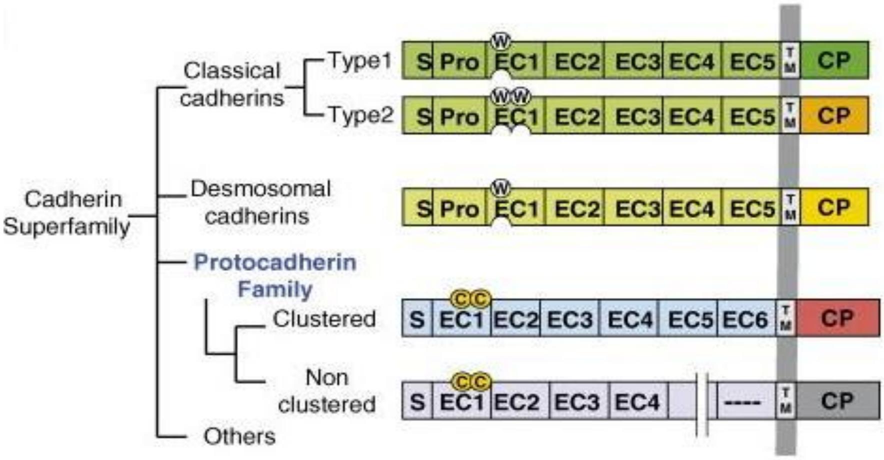


Figure 38. Schematic representation of members of the cadherin family. Some cadherins have a prodomain that is removed by a furin protease.

Classical cadherins, comprising six type I and thirteen type II cadherins, are structurally characterized by five extracellular domains (EC1-EC5), by a single pass transmembrane region and by a cytoplasmic tail which interacts, through the binding with intracellular molecules belonging to catenins superfamily, with the actin cytoskeleton (Figure 39).[5] They all share a similar primary sequence.

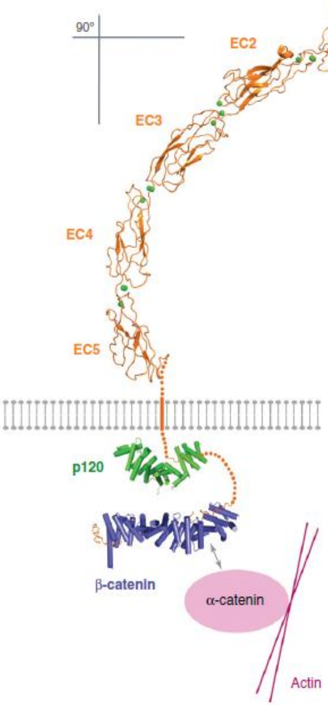


Figure 39. Overall architecture of classical cadherins.

In the adherens junctions, classical cadherins form *trans* complexes by bridging the intercellular space via their ectodomains. They project from opposing cell surfaces and form adhesive homodimers by interacting via their EC1 domains. The binding between cadherins and the actin cytoskeleton allows the structural stabilization of adherens junctions and promotes the regulation of cell morphology and motility.[13]

In addition to trans dimerization, the adhesion is strengthened by the lateral or *cis* association of cadherin ectodomains extending from the same cell surface. With respect to the trans dimer, the cis dimer structures appear to involve a different portion of EC1 that interacts with EC2 of a neighboring molecule emerging from the surface of the same cell.[14]

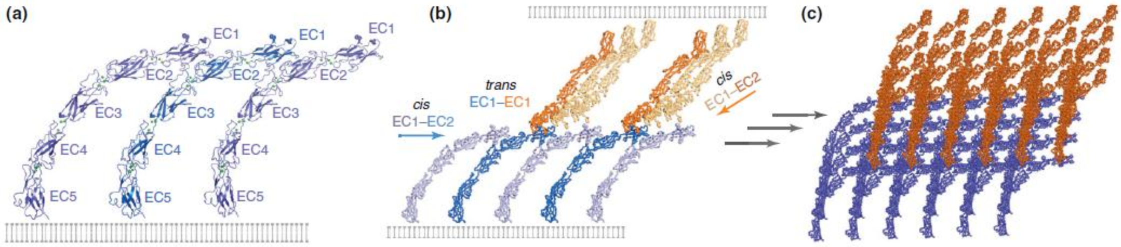


Figure 40. Proposed structural architecture of adherens junctions by X-ray dimer structures a) cis interaction, b) cis and trans interaction and c) the final net of cis and trans organized interactions.

On the basis of the recently published crystallographic structures of the whole E- and N-cadherin ectodomain dimers,[14] a model of adherent junction architecture has been proposed (Figure 40). According to the X-ray model, each cadherin monomer could simultaneously engage two molecules for lateral association and one for trans dimerization, resulting in an ordered two-dimensional layer that could represent the structural basis of the intercellular junction adhesion. Several experimental data[14,15] showed that lateral binding is weaker compared to trans homodimerization and it appears to exert a supporting role in cadherin-mediated adhesion[16] with respect to the primary function of the trans interface.

Among the type I classical cadherin subfamily, epithelial (E)-cadherin and neuronal (N)-cadherin have been the subject of my research activities. E-cadherin, expressed by epithelial cells, is regarded as the prototypical example of calcium-dependent homophilic cellular adhesion. N-cadherin, present in neural adherent junctions, also promotes cell migration during tissue morphogenesis.[17] Moreover, as I will discuss in the next section, both receptors have a central role during the progression of some types of cancer and are valuable targets for diagnostic and therapeutic applications.

## 5.3 ROLE OF N- AND E-CADHERINS IN CANCER

A number of physiologic processes influence the biology of adherens junctions. During development, for instance, the strength of inter-cellular adhesion may be modulated very rapidly in response to stimuli provided by growth factors and other molecules, without changes in the junctional complexes involved. Conversely, cellular differentiation and changes regarding the cellular transitions from a quiescent to a migratory state may induce and be induced by gross alterations in adherens junction assembly. An example is represented by the epithelial-mesenchymal transition (EMT), characterized as a phenotypic transition able to transform a quiescent epithelial cell in a highly motile and invasive cell. During cancer progression, in particular, epithelial cells undergoing EMT acquire the ability to migrate in a directional way, dissociating one from each other. These characteristics well explain the invasiveness and the ability to metastasize typical of malignant cancer cells. The reduction in E-cadherin expression appears to be the most important event during EMT. Cadherins, in fact, are expressed differentially during embryonic development and adult life. If E-cadherin is expressed predominantly by resting epithelia and is normally considered one of the principal suppressors of tumor invasiveness, N-cadherin is conversely present in the nervous system, smooth muscle cells, fibroblasts and endothelial cells but it is also de novo and aberrantly expressed by some human solid tumors (i.e. breast, prostate, thyroid and bladder cancer). Transcriptional repression of E-cadherin is considered one of the main events during neoplastic progression. Thus, during EMT E-cadherin is down-regulated and N-cadherin is de novo expressed at the same time, in a process called cadherin switching.[13,18,19]

Cadherin switching plays a pivotal role during neoplastic progression, and may arise contextually to tumor onset or when cancer cells change their phenotype from an epithelial to a mesenchymal one. Furthermore, the proinvasive action of N-cadherin persists even in the presence of E-cadherin. Forced expression of N-cadherin in breast cancer cell lines expressing E-cadherin does not reduces the expression of E-cadherin, while rendering the cells highly invasive and malignant. Conversely, exogenous expression of E-cadherin into a breast cancer cell line expressing N-cadherin does not impairs either its expression or invasive push. Moreover, human breast cancer metastases expressing N-cadherin co-express, in the murine model, both E- and N-cadherin in different anatomical sites. This suggests that N-cadherin may promote an increase in invasiveness and metastatic potential even if E-cadherin is expressed.[20]

The proinvasive activity of N-cadherin is reinforced by its functional interaction with Fibroblast Growth Factor (FGF) Receptor I (FGFR-I) on the cell surface. This interaction results primarily in the promotion of axonal growth. Several human cancer cell lines co-express N-cadherin and FGFR-I displaying, upon FGF stimulation, a high phosphorylation of cellular mediators involved in MAPK/ERK pathway. This ultimately leads to secretion of matrix metalloproteases and increase in invasiveness. This evidence suggests that the crosstalk between N-cadherin and

FGFR-I plays a pivotal role in the metastatic cascade in which cancer cells take advantage of N-cadherin in order to extravasate to surrounding tissues.[21]

E-cadherin is considered a repressor for the majority of carcinomas. However, it has been shown that in epithelial ovarian cancer (EOC) E-cadherin persists during tumor progression. E- and N-cadherins can be co-expressed in some advanced-stage EOCs, leading to the conclusion that E-cadherin expression and homophilic interaction contributes to the proliferation of EOC cells.[22]

## 5.4 STRUCTURE AND MECHANISM OF CADHERIN BINDING

The X-ray dimer structures of the whole EC1-EC5 ectodomain of E- and N-cadherins, published in 2011,[14] revealed a common interface underlying the homophilic binding. These structures, which are consistent with other structures of adhesive type I and type II ectodomain fragments, reveal a ‘‘strand swap’’ *trans* interface in which the EC1 N-terminal strand formed by residues DWVIPP (the *adhesion arm*) of each paired cadherin exchanges with that of the partner molecule. In order to form adhesive contacts, the adhesion arm of a cadherin molecule has to position the Trp side chain into the corresponding acceptor pocket in the EC1 of another molecule, located on the opposing cell surface, and form the swap dimer (Figure 41).

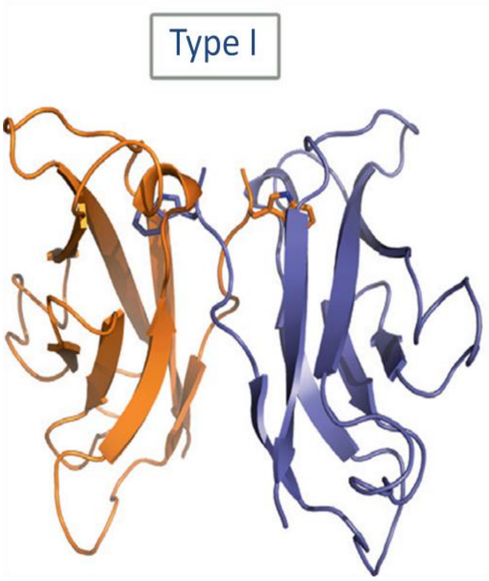


Figure 41. Swap dimer of classical cadherins formed by exchanging the monomers N-terminal DWVIPP β-strands. The two monomers come from different cells.

In this exchanging mechanism, the so-called 3D domain swapping,[23] the adhesive arm is first docked into its pocket and the monomer is in a *closed*, inactive form. The monomer then undergoes a conformational change leading to an *open*, active state with the adhesive arm exposed to the solvent and the swapping domain can occur (Figure 42).

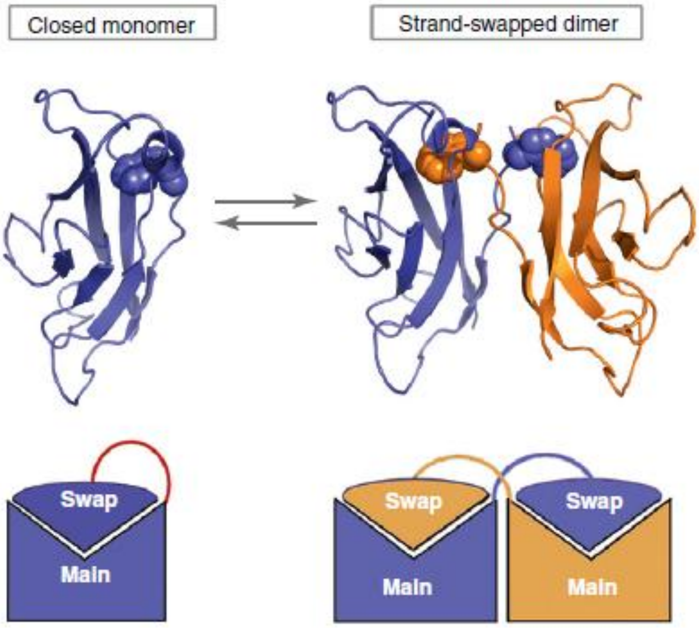


Figure 42. The adhesive binding mechanism of classical cadherins as an example of the 3D domain swapping.

While members of the same subfamily can form homo- and heterodimers, it has also been shown[24] that different cadherins subfamilies do not adhere to each other, suggesting a high degree of specificity in the adhesion process. In fact, the dimerization interface found in association with the crystal structures of type-II cadherins appears to be substantially different from that observed for type I cadherins, the former involving a larger portion of the EC1 domain. Since the *cis* binding interface, involving cadherins from the same cell, only accounts for a negligible amount to the overall adhesive binding affinity[15], cadherins specificity seems to be primarily modulated by the differences in strand-swapping interface.

For E-cadherin, multiple X-ray structures showing the extra-cellular domain of the dimerized complexes[25,26], various site-directed mutagenesis and electron microscopy studies[27,28] had highlighted the importance of this trans-swapped adhesive interface. On the other hand, N-cadherin was until the 2011 X-ray structure thought to dimerize forming a non-swapped complex, described by an X-ray structure dating back 1995 (Figure 43).[29]

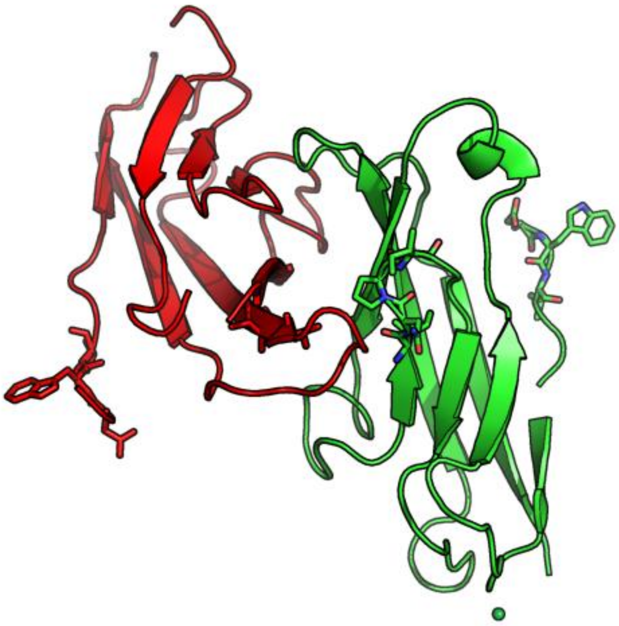


Figure 43. N-cadherin trans-dimer of 1995(1nch pdb), in which residues 53-55 and 79-81 form the adhesive interface.[29]

In this first trans dimer structure, N-cadherin EC1 monomers, supposedly coming from different cells, bind each other using the interface comprising residues 53-55 (INP) and 79-81 (HAV). This arrangement is characterized by the formation of a trans dimer with an antiparallel orientation of monomers. In this model, the N-terminal DWVIPP sequence is in turn thought to engage in a cis interaction with another cadherin coming from the same cell (Figure 44, same color), thus leading to a completely different cell-cell adhesion structure, which is depicted in Figure 44.

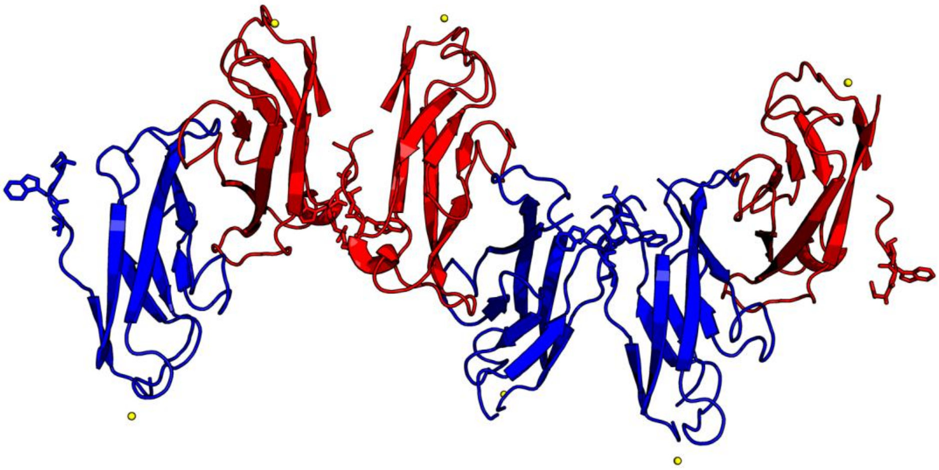


Figure 44. Model for the dimerization of N-cadherin based on X-ray dimer of 1995, in which monomers from the same cells have the same colors. Monomers coming from different cells interact through residues HAV (79-81) and INP (53-55).

Regarding the mechanism of the 3D domain swapping, many doubts still remain on the path leading to the cadherin dimerized structures. Various biophysical studies, from single-molecule FRET measurements,[30] to NMR relaxation experiments,[31] have been performed in order to characterize the full molecular mechanism of cadherin binding. Yet, it's still unclear whether cadherins dimerize through an induced-fit mechanism, with an intermediate, the so-called X-dimer (Figure 45), acting as the encounter complex that lowers the activation energy required for strand-swapping to occur, or rather the dimerization takes place through a selected-fit mechanism involving a conformational selection in which both monomers have to adopt an open, and therefore active, form prior to the dimerization. (Figure 46).

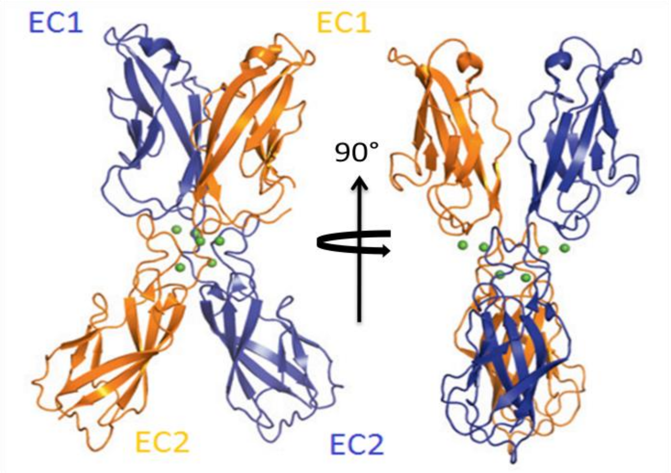


Figure 45. Front (left) and side view (right) of EC1-EC2 swap-impaired E-cadherin mutant K14E in the X-dimer form (pdb code: 3lne)[32].

In the induced fit hypothesis, regions of the conformational space virtually inaccessible to the monomer become reachable because of the encounter complex, which induces the conformational change, that is the opening of the arm, in both monomers.

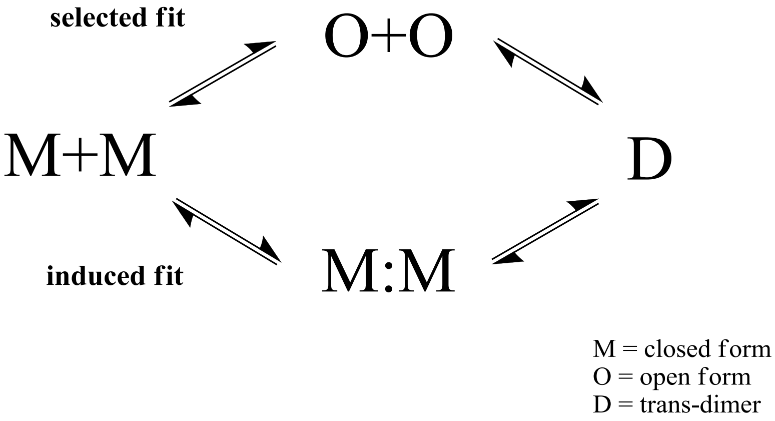


Figure 46. Proposed mechanisms for the dimerization of classical cadherins, selected fit (up) and induced fit (down) mechanisms. M:M stands for the encounter complex, the X-dimer.

Single-molecule FRET and atomic force measurements[30] on the full ectodomain have identified a Trp2 independent, $Ca^{2+}$ dependent, weak encounter complex, thus leading to the existence of an induced fit mechanism. However, subsequent crystallographic studies[14] have shown how E-cadherin mutants with mutations preventing the formation of the X-dimer - for example K14E introducing a charge repulsion with Asp138 - fully dimerize forming a trans-dimer. The authors concluded that the X-dimer configuration is a kinetically important intermediate in the dimerization of classical cadherins, though not strictly required. In addition, NMR measurements on the isolated EC1 of a type-II classical cadherin (cadherin 8) identified a small percentage of monomers with exposed Trp2 residues and no encounter complex.[31] This last result and the crystallographic data on mutated E-cadherins enable the possibility of dimerization through a selected fit mechanism.

## 5.5 TYPE I CLASSICAL CADHERIN ANTAGONISTS

Despite a growing interest in the field, the rational design of small ligands targeting cadherins protein-protein interactions (PPIs) is still in a very early stage.

Based on the structural model depicted in Figure 44, the first attempt was to block the supposed trans interface characterized by the His79-Ala80-Val81 (HAV) and the Ile53-Asn54-Pro55 (INP) sequences. So, libraries of disulfide-linked cyclic peptides based on HAV or INP sequences were synthesized.[21] Some of the peptides were shown to be able to modulate the outgrowth of neurite expressing N-cadherin either in “antagonistic” or “agonistic” modality. Among them, the antagonist peptide N-Ac-CHAVC-$NH_2$ (ADH-1 or Exherin$^{TM}$, Figure 47) containing the HAV motif, was shown to disrupt endothelial cell adhesion, induce apoptosis and inhibit angiogenesis in millimolar (mM) concentrations with favorable results in animal models of prostate and pancreas cancer, and of melanoma.[33–35]

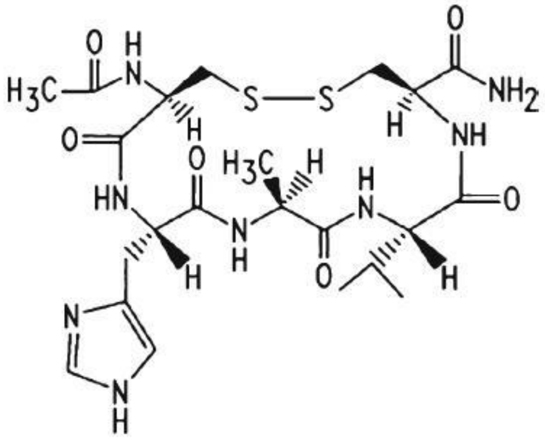
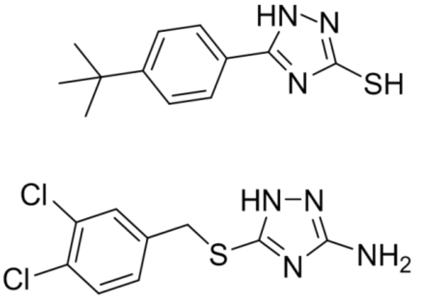


Figure 47. 2D structure of the the cyclic peptide ADH-1.

Thus, it was promoted to phase I clinical investigation in patients with advanced solid tumors which express N-cadherin, showing good tolerability in high doses, and some encouraging results suggestive of a potential therapeutic value.[7,36,37] Libraries of small peptide or non-peptide compounds were explored by a virtual screening protocol as possible mimics of HAV and related sequences, and some compounds were shown to inhibit the N-cadherin-mediated neurite outgrowth and cell adhesion (Figure 48).[38,39]



25

35

Figure 48. Most active compounds resulting from virtual screening of mimics of ADH-1.[38]

Phage display technology was employed to screen libraries of 12mer peptides against chimeric proteins composed of the human N-cadherin or E-cadherin ectodomains fused to the Fc fragment of human immunoglobulin G1.[40] All the isolated clones contained a Trp residue in position 2, while great variability existed throughout the other positions in the sequence. A linear peptide (H-SWELYYPLRANL-$NH_2$) was reproduced by synthesis and shown to inhibit the adhesion of human breast cancer cells expressing E- and N-cadherin in a mM range.

## 5.6 BIBLIOGRAPHY

(1) Arkin, M. R.; Wells, J. *Nat. Rev. Drug Discov.* **2004**, *3*, 301–317.

(2) Wells, J.; McClendon, C. L. *Nature* **2007**, *450*, 1001–1009.

(3) Berx, G.; van Roy, F. *Cold Spring Harb. Perspect. Biol.* **2009**, *1*, a003129.

(4) Leckband, D.; Sivasankar, S. *Curr. Opin. Cell Biol.* **2012**, *24*, 620–627.

(5) Brasch, J.; Harrison, O. J.; Honig, B.; Shapiro, L. *Trends Cell Biol.* **2012**, *22*, 299–310.

(6) Blaschuk, O. W.; Devemy, E. *Eur. J. Pharmacol.* **2009**, *625*, 195–198.

(7) Blaschuk, O. W. *Cell Tissue Res.* **2012**, *348*, 309–313.

(8) Adams, C. L.; Chen, Y. T.; Smith, S. J.; Nelson, W. J. *J. Cell Biol.* **1998**, *142*, 1105–1019.

(9) Garrod, D. R.; Merritt, A. J.; Nie, Z. *Curr. Opin. Cell Biol.* **2002**, *14*, 537–545.

(10) Takeichi, M. *Science* **1991**, *251*, 1451–1455.

(11) Boggon, T. J.; Murray, J.; Chappuis-Flament, S.; Wong, E.; Gumbiner, B. M.; Shapiro, L. *Science* **2002**, *296*, 1308–1313.

(12) Pokutta, S.; Herrenknecht, K.; Kemler, R.; Engel, J. *Eur. J. Biochem.* **1994**, *223*, 1019–1026.

(13) Gumbiner, B. M. *J. Cell Biol.* **2000**, *148*, 399–404.

(14) Harrison, O. J.; Jin, X.; Hong, S.; Bahna, F.; Ahlsen, G.; Brasch, J.; Wu, Y.; Vendome, J.; Felsovalyi, K.; Hampton, C. M.; Troyanovsky, R. B.; Ben-Shaul, A.; Frank, J.; Troyanovsky, S. M.; Shapiro, L.; Honig, B.; Sergey, M. *Structure* **2011**, *19*, 244–256.

(15) Brasch, J.; Harrison, O. J.; Honig, B.; Shapiro, L. *Trends Cell Biol.* **2012**, *22*, 299–310.

(16) Wu, Y.; Jin, X.; Harrison, O.; Shapiro, L.; Honig, B. H.; Ben-Shaul, A. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 17592–17597.

(17) Halbleib, J. M.; Nelson, W. J. *Genes Dev.* **2006**, *20*, 3199–3214.

(18) Cavallaro, U.; Liebner, S.; Dejana, E. *Exp. Cell Res.* **2006**, *312*, 659–667.

(19) Jeanes, A.; Gottardi, C. J.; Yap, A. S. *Oncogene* **2008**, *27*, 6920–6929.

(20) Nagi, C.; Guttman, M.; Jaffer, S.; Qiao, R.; Keren, R.; Triana, A.; Li, M.; Godbold, J.; Bleiweiss, I. J.; Hazan, R. B. *Breast Cancer Res. Treat.* **2005**, *94*, 225–235.

(21) Williams, G.; Williams, E.-J.; Doherty, P. *J. Biol. Chem.* **2002**, *277*, 4361–4367.

(22) De Santis, G.; Miotti, S.; Mazzi, M.; Canevari, S.; Tomassetti, A. *Oncogene* **2009**, *28*, 1206–1217.

(23) Gronenborn, A. M. *Curr. Opin. Struct. Biol.* **2009**, *19*, 39–49.

(24) Chen, C. P.; Posy, S.; Ben-Shaul, A.; Shapiro, L.; Honig, B. H. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 8531–8536.

(25) Pertz, O.; Bozic, D.; Koch, A. W.; Fauser, C.; Brancaccio, A.; Engel, J. *EMBO J.* **1999**, *18*, 1738–1747.

(26) Parisini, E.; Higgins, J. M. G.; Liu, J.; Brenner, M. B.; Wang, J. *J. Mol. Biol.* **2007**, *373*, 401–411.

(27) Meng, W.; Takeichi, M. *Cold Spring Harb. Perspect. Biol.* **2009**, *1*, a002899.

(28) Shapiro, L.; Weis, W. I. *Cold Spring Harb. Perspect. Biol.* **2009**, *1*, a003053.

(29) Shapiro, L.; Fannon, A. M.; Kwong, P. D.; Thompson, A.; Lehmann, M. S.; Grübel, G.; Legrand, J. F.; Als-Nielsen, J.; Colman, D. R.; Hendrickson, W. A. *Nature* **1995**, *374*, 327–337.

(30) Sivasankar, S.; Zhang, Y.; Nelson, W. J.; Chu, S. *Structure* **2009**, *17*, 1075–1081.

(31) Miloushev, V. Z.; Bahna, F.; Ciatto, C.; Ahlsen, G.; Honig, B.; Shapiro, L.; Palmer, A. G. *Structure* **2008**, *16*, 1195–1205.

(32) Harrison, O. J.; Bahna, F.; Katsamba, P. S.; Jin, X.; Brasch, J.; Vendome, J.; Ahlsen, G.; Carroll, K. J.; Price, S. R.; Honig, B.; Shapiro, L. *Nat. Struct. Mol. Biol.* **2010**, *17*, 348–357.

(33) Li, H.; Price, D. K.; Figg, W. D. *Anticancer. Drugs* **2007**, *18*, 563–568.

(34) Shintani, Y.; Fukumoto, Y.; Chaika, N.; Grandgenett, P. M.; Hollingsworth, M. A.; Wheelock, M. J.; Johnson, K. R. *Int. J. Cancer* **2008**, *122*, 71–77.

(35) Augustine, C. K.; Yoshimoto, Y.; Gupta, M.; Zipfel, P. A.; Selim, M. A.; Febbo, P.; Pendergast, A. M.; Peters, W. P.; Tyler, D. S. *Cancer Res.* **2008**, *68*, 3777–3784.

(36) Perotti, A.; Sessa, C.; Mancuso, A.; Noberasco, C.; Cresta, S.; Locatelli, A.; Carcangiu, M. L.; Passera, K.; Braghetti, A.; Scaramuzza, D.; Zanaboni, F.; Fasolo, A.; Capri, G.; Miani, M.; Peters, W. P.; Gianni, L. *Ann. Oncol.* **2009**, *20*, 741–745.

(37) Yarom, N.; Stewart, D.; Malik, R.; Wells, J.; Avruch, L.; Jonker, D. J. *Curr. Clin. Pharmacol.* **2013**, *8*, 81–88.

(38) Gour, B. J.; Blaschuk, O. W.; Ali, A.; Ni, F.; Chen, Z.; Michaud, S. D.; Shoameng, W.; Hu, Z. Peptidomimetic modulators of cell adhesion. US7446120 B2, 2008.

(39) Burden-Gulley, S. M.; Gates, T. J.; Craig, S. E. L.; Lou, S. F.; Oblander, S.; Howell, S.; Gupta, M.; Brady-Kalnay, S. M. *Peptides* **2009**, *30*, 2380–2387.

(40) Devemy, E.; Blaschuk, O. W. *Peptides* **2009**, *30*, 1539–1547.

# Chapter 6: Characterization of E- and N-cadherin binding interface

## 6.1 INTRODUCTION

E- and N-cadherin show a very high resemblance in the primary sequence. After 1D alignment[1] (using a Smith-Waterman algorithm[2]) of the first two Extra-Cellular (EC) domains, similarity sums up to 80 %, with 56 % identical amino acids (Figure 49).

```
               10        20        30        40        50        60        70        80
3Q2V:  DWVIPPISCPENEKGEFPKNLVQIKSNRDKETKVFYSITGQGADKPPVGVFIIERETGWLKVTQPLDREAIAKYILYSHA
       :::::::. ::: .: ::..::.:.:.:::. .. ::.:: :::.::.:.:::.  .: :.::.:::: ::.. : .::
3Q2W:  DWVIPPINLPENSRGPFPQELVRIRSDRDKNLSLRYSVTGPGADQPPTGIFIINPISGQLSVTKPLDRELIARFHLRAHA
               10        20        30        40        50        60        70        80

               90       100       110       120       130       140       150       160
3Q2V:  VSSNGEAVEDPMEIVITVTDQNDNRPEFTQEVFEGSVAEGAVPGTSVMKVSATDADDDVNTYNAAIAYTIVSQDPELPHK
       :. ::. ::.:..:::.: :.::::::: ..:..::: ::. ::: :: :.: ::::  :. :. . : :.:: :  :
3Q2W:  VDINGNQVENPIDIVINVIDMNDNRPEFLHQVWNGSVPEGSKPGTYVMTVTAIDADDP-NALNGMLRYRILSQAPSTPSP
               90       100       110       120       130        140       150

              170       180       190          200       210
3Q2V:  NMFTVNRDTGVISVLTSGLDRESYPTYTLVVQAADLQGE---GLSTTAKAVITVKDIND
       ::::.: .:: : ....:::::.   :::..::.:..:.   :::.:: ::::: :.
3Q2W:  NMFTINNETGDIITVAAGLDREKVQQYTLIIQATDMEGNPTYGLSNTATAVITVTDV
     160       170       180       190       200       210
```

Figure 49. 1D sequence alignment[f] using a Smith-Waterman algorithm for E-cadherin (3q2v) and N-cadherin (3q2w) EC1-EC2 domains.

The correspondence in the primary structure is also conserved in the secondary structure. In fact, EC1-EC2 Cα alignment of the X-ray structures[3] of N-cadherin (pdb: 3q2w) and E-cadherin (pdb: 3q2v) shows a striking almost perfect superposition of the two 3D structures (Figure 50, left). What is more, all the secondary structure elements are shared among the two types of classical cadherins, with one α-helix per EC domain and a large number of β-strands (Figure 50, right).

[f] http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=compare

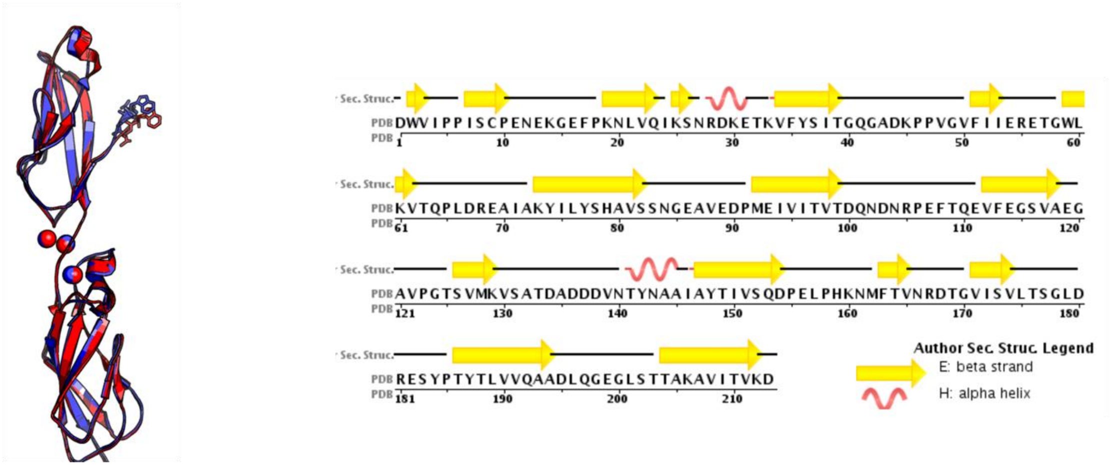
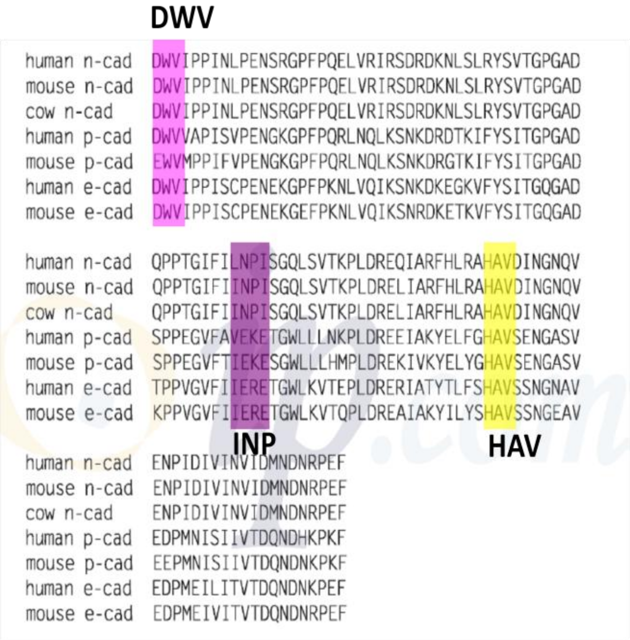

Sec. Struc.
PDB DWVIPPISCPENEKGEFPKNLVQIKSNRDKETKVFYSITGQGADKPPVGVFIIERETGWL
PDB 1 10 20 30 40 50 60
Sec. Struc.
PDB KVTQPLDREAIAKYILYSHAVSSNGEAVEDPMEIVITVTDQNDNRPEFTQEVFEGSVAEG
PDB 61 70 80 90 100 110 120
Sec. Struc.
PDB AVPGTSVMKVSATDADDDVNTYNAAIAYTIVSQDPELPHKNMFTVNRDTGVISVLTSGLD
PDB 121 130 140 150 160 170 180
Sec. Struc.
PDB RESYPTYTLVVQAADLQGEGLSTTAKAVITVKD
PDB 181 190 200 210
Author Sec. Struc. Legend
E: beta strand
H: alpha helix

Figure 50. Left: EC1-EC2 Cα alignment between N-cadherin (3q2w, red) and E-cadherin (3q2v, blue). RMSD is 0.615 Å. Right: secondary structure elements for EC1-EC2 N-cadherin (pdb: 3q2w), which are entirely shared with E-cadherin (not shown).

However, the crystallographic data of the E- and N-cadherin homodimers have shown a possible different dimerization interface for N-cadherin (pdb code 1nch).[4] As already discussed in the previous chapter, there is a first non-swapped dimer of EC1 N-cadherin domains (1995, 1nch.pdb), and a new X-ray crystal structure of the whole N-cadherin ectodomain (2011, 3q2w.pdb) that dimerize through a swapping of the N-terminal β-strand similar to that observed in the X-ray structure of E-cadherin dimers. The INPI sequence identified by the non-swapped N-cadherin dimer is selectively present only in N-cadherin, while the HAV tripeptide, which is supposed to interact with the INPI sequence at the interface and included into the ADH-1 antagonist,[5] is highly conserved among various mammals type I cadherins (Figure 51).



```
human n-cad   DWVIPPINLPENSRGPFPQELVRIRSDRDKNLSLRYSVTGPGAD
mouse n-cad   DWVIPPINLPENSRGPFPQELVRIRSDRDKNLSLRYSVTGPGAD
cow n-cad     DWVIPPINLPENSRGPFPQELVRIRSDRDKNLSLRYSVTGPGAD
human p-cad   DWVVAPISVPENGKGPFPQRLNQLKSNKDRDTKIFYSITGPGAD
mouse p-cad   EWVMPPIFVPENGKGPFPQRLNQLKSNKDRGTKIFYSITGPGAD
human e-cad   DWVIPPISCPENEKGPFPKNLVQIKSNKDKEGKVFYSITGQGAD
mouse e-cad   DWVIPPISCPENEKGEFPKNLVQIKSNRDKETKVFYSITGQGAD

human n-cad   QPPTGIFILNPISGQLSVTKPLDREQIARFHLRAHAVDINGNQV
mouse n-cad   QPPTGIFIINPISGQLSVTKPLDRELIARFHLRAHAVDINGNQV
cow n-cad     QPPTGIFIINPISGQLSVTKPLDRELIARFHLRAHAVDINGNQV
human p-cad   SPPEGVFAVEKETGWLLLNKPLDREEIAKYELFGHAVSENGASV
mouse p-cad   SPPEGVFTIEKESGWLLLHMPLDREKIVKYELYGHAVSENGASV
human e-cad   TPPVGVFIIERETGWLKVTEPLDRERIATYTLFSHAVSSNGNAV
mouse e-cad   KPPVGVFIIERETGWLKVTQPLDREAIAKYILYSHAVSSNGEAV

human n-cad   ENPIDIVINVIDMNDNRPEF
mouse n-cad   ENPIDIVINVIDMNDNRPEF
cow n-cad     ENPIDIVINVIDMNDNRPEF
human p-cad   EDPMNISIIVTDQNDHKPKF
mouse p-cad   EEPMNISIIVTDQNDNKPKF
human e-cad   EDPMEILITVTDQNDNKPEF
mouse e-cad   EDPMEIVITVTDQNDNRPEF
```

Figure 51. 1D alignment of the primary sequences of different type I classical cadherins.

Considering the adhesive arm of the swap dimer interface of both crystal structures of E- and N-cadherin, i.e. the N-terminal DWVIPP sequence, we observe a very high degree of similarity among type I classical cadherins. In particular, Trp in position 2 is conserved in all six type I classical cadherins.[6]

In order to characterize both the N- and E-cadherin homophilic interfaces, different computational techniques were used. Binding sites prediction tools were employed to deduct the energy hot spots in the new crystal structures of N-cadherin (pdb code 3q2w) and E-cadherin (pdb code 3q2v). Computational alanine scanning was performed on N-cadherin, using both the 1995 dimer model (pdb code 1nch) and the new dimer structure (pdb code 3q2w). Finally, the dynamic behavior of N and E-cadherin dimers was analyzed by performing Molecular Dynamics (MD) simulations using the new structures 3q2v and 3q2w for E- and N cadherin, respectively.

## 6.2 BINDING SITE PREDICTION TOOLS

Our first task was to try and predict which are the interaction hot spots in N-cadherin and E-cadherin homophilic binding. Site prediction tools can be used for this purpose to locate the preferred binding sites in a protein-protein interaction, using only the 3D structure of a monomeric protein.

For the N-cadherin, we selected and used two binding site prediction tools (SiteMap, which is part of the Schrödinger suite[7] and QSiteFinder, a web server[g,8]). In these tools, the algorithm by which a site is identified and ranked with respect to the others, works by dividing the protein space in a point grid and evaluating the interaction energy between "probe" molecules, i.e. having hydrophobic or hydrophilic properties, and each point of the grid.[9] While SiteMap employs a variety of probes each having specific chemical properties, QSiteFinder specializes in finding hydrophobic sites, by using a probe which simulates a methyl group. A recent review by Nussinov and coworkers[10] on available Protein-Protein Interactions (PPI) prediction tools contains detailed information.

Using the EC1-EC5 N-cadherin monomer structure as input (3q2w.pdb), both the tools identified as top ranked and most important binding site the hydrophobic pocket onto which the adhesive arm docks the side chain of Trp2, formed by residues Ile24, Ser26, Tyr36, Ala78, Ala80, Asn90 and Ile92, while simultaneously failing at identifying the putative interface HAV-INPI (Figure 52).

[g] http://www.modelling.leeds.ac.uk/qsitefinder/

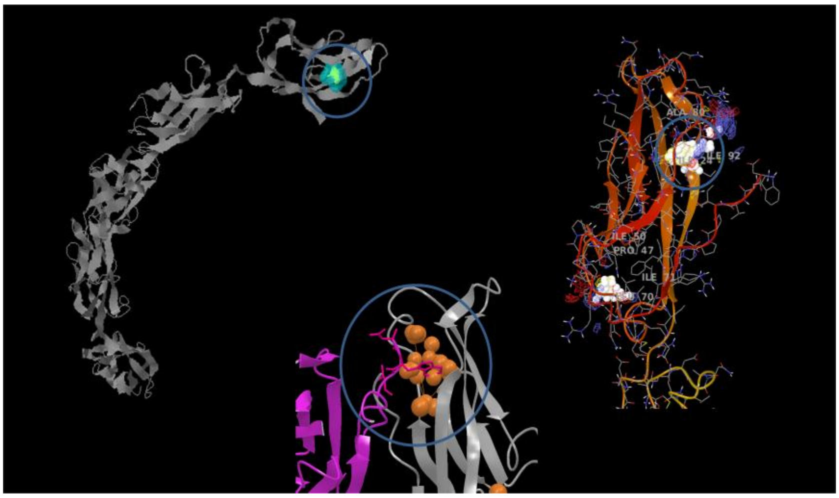


Figure 52. Top ranked binding site identified for N-cadherin by QSiteFinder (top left) and SiteMap (top right). In both cases the top ranked pocket cavity is the one onto which the adhesive arm docks Trp2 (bottom).

SiteMap ranks each binding site according to a score function called SiteScore, in which a weighted sum of calculated properties is performed. The properties include *size,* the number of grid points that make up the site, *en* the degrees of enclosure by the protein and *ex*, exposure to solvent, and the hydrophobic (*phobic)* and hydrophilic (*philic*) character of the site (Table 5).

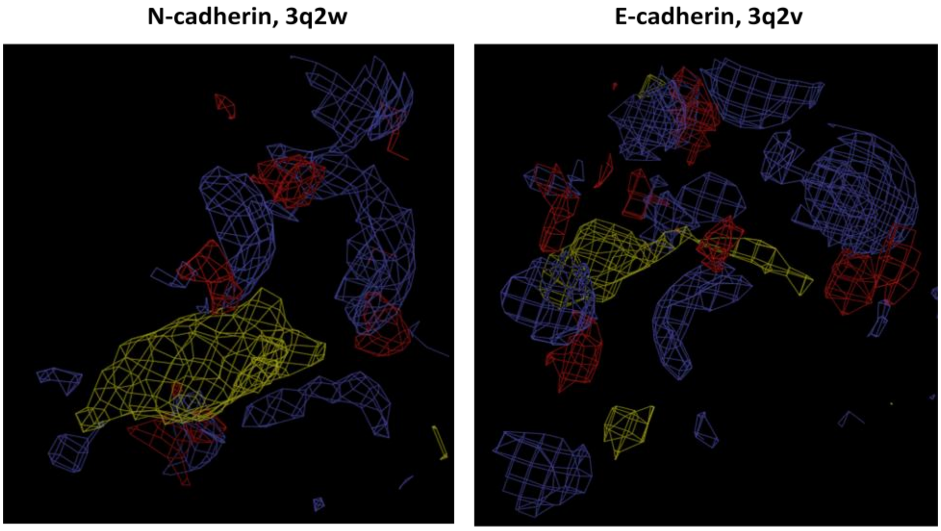
| Title | SiteScore | size | ex | en | phobic | philic |
|---|---|---|---|---|---|---|
| site 1 | 0.91 | 47 | 0.64 | 0.68 | 1.94 | 0.48 |
| site 2 | 0.71 | 25 | 0.52 | 0.71 | 1.82 | 0.70 |
| site 3 | 0.65 | 33 | 0.66 | 0.65 | 0.35 | 1.36 |
| site 4 | 0.61 | 29 | 0.68 | 0.62 | 0.48 | 1.03 |
| site 5 | 0.59 | 30 | 0.80 | 0.54 | 0.22 | 0.89 |

Table 5. Top ranked binding sites of N-cadherin as predicted by SiteMap. Site 1 represents the pocket to which Trp2 binds in the trans-swapped dimer (pdb: 3q2w).

The only binding site with a SiteScore value above 0.8 (Table 5), that according to the authors distinguishes drug-binding and non drug binding sites, correspond to the Trp2 binding pocket. In addition, the second best site is still located at EC1 but far from the INPI-HAV residues while the other putative binding sites are found between EC2 and EC5.

These results suggest that the Trp2-accepting pocket and the corresponding adhesion arm could be the hot spots of the N-cadherin homophilic interaction.

As described in the introduction, N- and E-cadherin share a great similarity both in the primary and the secondary structure. However, small differences in the sequence are present. Some of them are indeed localized in the Trp2 binding pocket and make up for a slight reduction in the hydrophobicity of the E-cadherin binding pocket. SiteMap analysis performed for E-cadherin identified the same top ranked binding site of N-cadherin with a SiteScore value of 0.87, slightly lower than the one obtained for N-cadherin. The Trp2 binding site showed in fact a less hydrophobic character (Figure 53, yellow isosurface).



N-cadherin, 3q2w

E-cadherin, 3q2v

Figure 53. SiteMap results visualized for N-cadherin (left) and E-cadherin(right). In yellow, red and blue are shown the hydrophobic, H-bond acceptor and H-bond donor isosurfaces, respectively.

Residue mutations are indeed observed into the two binding sites: Asp27, Ala78, Asn90, Ile92 in the N-cadherin site are replaced by Asn27, Ser78, Asp90, Met92 in E-cadherin (Figure 54).

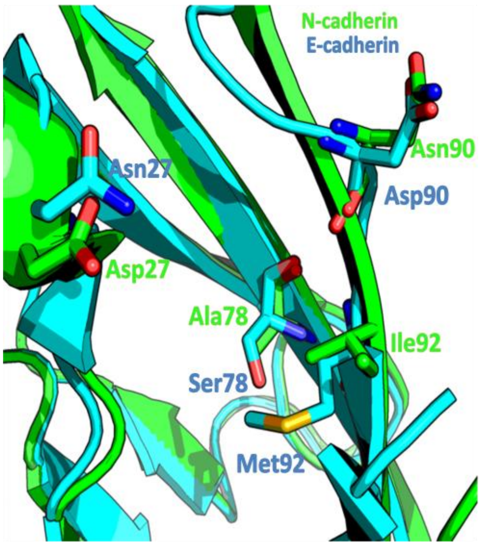


Figure 54. N-cadherin (green) and E-cadherin (light blue) EC1. Residues near the binding pocket which differs among the two cadherins are shown as sticks.

## 6.3 Computational alanine scanning

The prediction tools described in the preceding section estimate the most probable binding site without any knowledge of the actual 3D structure of the protein-protein dimer. In fact they require only the monomer protein as input.

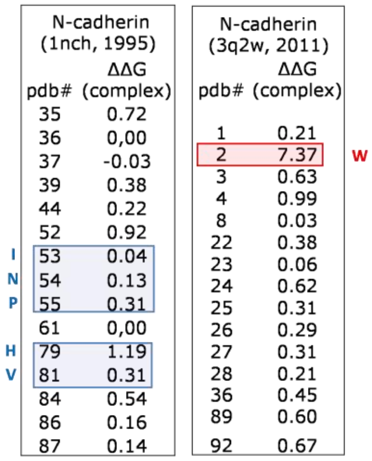
A different approach consists in exploiting the known putative dimer adhesion structures and in performing on them a computational alanine scanning.[11] By consecutively mutating each amino acid of the monomers to alanine and calculating the difference in the binding energy between the wild type dimer and the mutated one, the contribution of each residue to the total binding energy can be assessed. Here, we used the two available N-cadherin dimer structures, 1nch and 3q2w.

To perform computational alanine scanning, a wide selection of web servers exists, and we chose the Robetta Web Server,[h] developed by Baker and coworkers.[12] The web version of the alanine scanning algorithm differs from the one described in Chapter 2 as it does not use any molecular dynamics simulation, but solely relies on the three dimensional structure of a protein-protein complex. Then, sequentially, each amino acid is mutated to alanine and a simple free energy functional form is used to calculate the free energy of binding of both the wild type and the mutated complexes. The difference in $\Delta G_{binding}$ between the wild type and each mutated species is then computed:

$$\Delta\Delta G = (\Delta G_D^{WT} - \Delta G_A^{WT}\Delta G_B^{WT}) - (\Delta G_D^{m} - \Delta G_A^{m}\Delta G_B^{m}) \qquad (32)$$

where the subscripts $A$, $B$ and $D$ refer to the two monomers and the dimer, respectively, and the superscripts $WT$ and $m$ refer to the wild type and the mutated species. A value of $\Delta\Delta G_{binding}$ much higher than 1 kcal/mol is indicative of an important residue that, if mutated to alanine, greatly destabilizes the binding affinity between the two partners.

The computational alanine scanning results obtained for both N-cadherin dimers (1nch and 3q2w) are summarized in Table 6.



| | N-cadherin (1nch, 1995) | |
|---|---|---|
| | pdb# | ΔΔG (complex) |
| | 35 | 0.72 |
| | 36 | 0,00 |
| | 37 | -0.03 |
| | 39 | 0.38 |
| | 44 | 0.22 |
| | 52 | 0.92 |
| I | 53 | 0.04 |
| N | 54 | 0.13 |
| P | 55 | 0.31 |
| | 61 | 0,00 |
| H | 79 | 1.19 |
| V | 81 | 0.31 |
| | 84 | 0.54 |
| | 86 | 0.16 |
| | 87 | 0.14 |

| N-cadherin (3q2w, 2011) | | |
|---|---|---|
| pdb# | ΔΔG (complex) | |
| 1 | 0.21 | |
| 2 | 7.37 | W |
| 3 | 0.63 | |
| 4 | 0.99 | |
| 8 | 0.03 | |
| 22 | 0.38 | |
| 23 | 0.06 | |
| 24 | 0.62 | |
| 25 | 0.31 | |
| 26 | 0.29 | |
| 27 | 0.31 | |
| 28 | 0.21 | |
| 36 | 0.45 | |
| 89 | 0.60 | |
| 92 | 0.67 | |

Table 6. Computational alanine scanning results for N-cadherin performed on the 1995 3D dimer structure (1nch) and the 2011 3D structure (3q2w).

[h] http://robetta.bakerlab.org

Only the mutation of the Trp2 residue in the swap dimer interface has a pronounced effect on the overall free energy of binding. What is really interesting is that the mutation of the residues forming the supposed hot spots of interaction in the 1995 crystal structure (HAV-INPI), does not affect very much the $\Delta G_{binding}$. Only residue 79, if mutated, reaches a value slightly higher than 1 kcal/mol.

In conclusion, our in silico analysis, supported by several experimental data,[13] has shown that probably the first proposed binding interface for N-cadherin is not representative of a real mode of intercellular interaction, but could derive from crystallographic artifacts. As in this 1995 X-ray structure only the EC1 domains were crystallized, the consequent model developed to interpret the cell-cell trans and cis interactions, inevitably neglected the effect of all the other extracellular domains. We then focused our attention to the binding hot spot involving the Trp2, common to all type I classical cadherins.

## 6.4 MOLECULAR DYNAMICS SIMULATION OF E- AND N-CADHERIN DIMERS

In order to better analyze the relevant binding features of the adhesive interfaces, and to verify whether small changes in the E- and N-cadherin binding site could affect the swap-dimer interface, we then performed Molecular Dynamics (MD) simulations of 50 ns (AMBER 11,[14] T=300K, TIP3P[15] water model) starting from the EC1-EC2 fragment of the N- and E-cadherin X-ray dimer structures (3q2w.pdb and 3q2v.pdb, respectively).

During the simulations the two systems maintained the key crystallographic contacts of the DWVI adhesive sequence showing a nearly identical pattern of interactions, that can be summarized as follows (Figure 55):

1. the formation of an intermolecular salt bridge between the charged N-terminal amino group of Asp1 and the side chain carboxylate of Glu89

2. the anchoring of the Trp2 side chain into a hydrophobic pocket
3. the formation of a hydrogen bond between the indole moiety and the carbonyl group of Asn90 (N-cadherin) or Asp90 (E-cadherin)
4. the involvement of Val3-NH in a hydrogen bond with the carbonyl group of Arg25 (N-cadherin) or Lys25 (E-cadherin)
5. the formation of a hydrogen bond between the backbone carbonyl group of Asp1 and the Asp27-NH (N-cadherin) or Asn27-NH (E-cadherin) group.

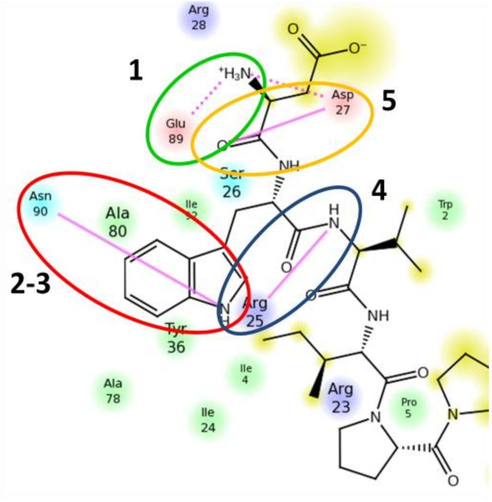


Figure 55. 2D map illustrating the contacts maintained during the MD simulation of E-and N-cadherin. Here only the N-cadherin binding pocket (with residues colored based on their properties and their distance from the adhesive arm) is shown.

In Table 7 the monitored crystallographic contacts are reported for both N- and E-cadherin, considering that each dimer has two EC1 domains interacting each other and acting as ligand, using the DWVI adhesive arm, and as receptor at the same time. Both systems keep the input crystallographic interactions of the DWVI sequence. The main difference in the interaction pattern observed for the DWVI motif in the two systems is only limited to a salt bridge formed between the Asp1-$NH_3^+$ group and the carboxyl group of Asp27 in N-cadherin binding site that in E-cadherin receptor is replaced by a hydrogen bond with the side chain of Asn27. The positively charged N-terminus of N-cadherin can in fact form two salt bridges at the same time with Glu89 and Asp27. On the contrary, E-cadherin position 27 is mutated to Asn, and no additional salt bridge can be formed.

| **Interaction (Ligand-Receptor Residues)** | **N-cadherin** | | **E-cadherin** | |
|---|---|---|---|---|
| | **$L_A/R_B$** | **$L_B/R_A$** | **$L_A/R_B$** | **$L_B/R_A$** |
| $Asp^1NH_3^+/C^{\delta}OO^-Glu^{89*}$ (1) | 88 | 98 | 100 | 100 |
| $Trp^2N^{\varepsilon 1}H\text{--}COAsn_{N\text{-}cadh}{}^{90}/COAsp_{E\text{-}cadh}{}^{90**}$ (3) | 99 | 99 | 98 | 98 |
| $Val^3NH\text{--}COArg_{N\text{-}cadh}{}^{25}/COLys_{E\text{-}cadh}{}^{25**}$ (4) | 98 | 96 | 99 | 99 |
| $Asp^1CO\text{--}NHAsp_{N\text{-}cadh}{}^{27}/NHAsn_{E\text{-}cadh}{}^{27**}$ (5) | 96 | 97 | 99 | 98 |
| $Asp^1NH_3^+/C^{\gamma}OO^-Asp_{N\text{-}cadh}{}^{27}/C^{\gamma}OAsn_{E\text{-}cadh}{}^{27}$* | 24 | 43 | 78 | 77 |

*distance between N and C < 4.0 Å,** distance between H and O < 2.5 Å

Table 7. Percentage of MD structures forming the X-ray interactions of the DWVI sequence observed in the E- and N-cadh swap dimers. $L_A$ and $L_B$ represent the DWVI sequence belonging to molecule A and B, respectively, while $R_A$ and $R_B$ the corresponding receptor pocket. To form the dimer, $L_A$ interacts with $R_B$ and $L_B$ with $R_A$.

## 6.5 Conclusions

The initial study performed only on the N-cadherin ectodomain in order to identify the most probable binding site for the homophilic trans interaction, has permitted us to verify the analogy in the binding sites for both E-cadherin and N-cadherin, discarding a previously supposed binding mode peculiar for only N-cadherin. The analysis of the most conserved contacts during the MD simulations has helped us to draw a complete map of the needed interactions in the trans-dimer structures. The information obtained was then used to develop putative inhibitors of the E-cadherin and N-cadherin homophilic interaction.

## 6.6 METHODS

### 6.6.1 QSiteFinder and SiteMap

QSiteFinder input was the pdb of the full ectodomain of N-cadherin (pdb code 32qw) while SiteMap was performed on both N-cadherin EC1-EC5 (pdb code 3q2w) and E-cadherin (pbd code 3q2v). EC1-EC5. The algorithm[16] for QSiteFinder is a slightly modified version of the protocol developed by Goodford.[9] Only an hydrophobic probe can be used to find putative binding sites.

SiteMap requires an optimized protein starting structure. As a consequence, the Protein Preparation Wizard, from the Maestro[17] Graphical User Interface, has been used to add hydrogen atoms, optimize hydrogen bonds and minimize the structures.

### 6.6.2 Robetta Web Server

N-cadherin complex 3q2w was stripped of all the Extra-Cellular domains from EC2 to EC5 in order for the results to be comparable to the N-cadherin complex 1nch, in which only the EC1 was crystallized. The mutated residues were selected to be from 1 to 99. Only one $Ca^{2+}$ ion at the end of EC1 was kept, in correspondence to 1nch.

The functional form of the free energy consists of a linear combination of a Lennard-Jones potential ($E_{LJ}$) to describe atomic packing interactions, an implicit solvation model ($G_{sol}$), an orientation-dependent hydrogen-bonding potential ($E_{HB}$) derived from high-resolution protein structures, statistical terms approximating the backbone-dependent amino acid-type and rotamer probabilities ($E_{\Phi\Psi}$), and an estimate of unfolded reference state energies ($E_{aa}^{ref}$), each relatively weighted (W):

$$\begin{aligned}\Delta G = {} & W_{attr}E_{LJattr} + W_{rep}E_{LJrep} + W_{HB(sc-bb)}E_{HB(sc-bb)} \\ & + W_{HB(sc-sc)}E_{HB(sc-sc)} + W_{sol}G_{sol} \\ & + W_{\Phi\Psi}E_{\Phi\Psi}(\mathrm{aa}) + \sum_{aa=1}^{20} n_{aa}E_{aa}^{ref}\end{aligned} \tag{33}$$

### 6.6.3 MD simulations

#### *6.6.3.1 Proteins preparation*

We built the EC1-EC2 dimer systems starting from the X-ray swap dimer structures of the E- and N-cadherin (pdb codes 3q2v and 3q2w, respectively). Each EC1-EC2 chain was truncated at residue number 218. Lys14 and Glu16 missing residues of E-cadherin chain A and Lys30 CD, CE and NZ missing atoms of N-cadherin chains were manually added. Three calcium ions $Ca^{2+}$ were kept at the interface of EC1-EC2 domains and one at the end of EC2 domain (Ca605 and Ca604 for E- and N-cadherin, respectively). All sugars and crystallographic waters were removed during the input preparation. In addition, for the E-cadherin dimer, two manganese ions each coordinated to Glu13 side chain have been removed. The two systems were then prepared using the Protein Preparation Wizard of the Maestro graphical user interface[17] by optimizing the orientation of hydrogen bonds and charge interactions, and predicting the protonation state of histidine, aspartic and glutamic acid and the tautomeric state of histidine, followed by a restrained minimization of the whole system (0.30 Å of RMSD on heavy atom) using the OPLSAA force field. The final refined structures were used to generate docking receptor grids and as input for Molecular Dynamic (MD) simulations.

#### *6.6.3.2 MD setup and calculation*

MD simulations were performed using the AMBER 11 package[14] with the ff10 force field.[18] Calcium ions were modeled on the basis of parameters reported by

Bradbrook[19] and histidine residues were set to HID (histidine with hydrogen on the delta nitrogen). The two systems were solvated in a cubic box with a 12 Å buffer by adding 48606 TIP3P waters for E-cadherin and 41527 for N-cadherin and $Na^+$ counterions were added to ensure electroneutrality.

To allow the systems to relax and release the strain due to crystal-packing effects, the two dimers were minimized keeping the complex fixed and just minimizing the positions of water and ions (with an harmonic restraint potential of force constant of k=10 kcal/molÅ$^2$), then the dimers were energy minimized restraining the position of relaxed waters and the ions (k=10 kcal/molÅ$^2$), and finally the entire systems were energy miminized unrestrained, by performing 2000 steps of steepest descent algorithm. Afterwards, the temperature of the system was slowly brought to the desired value of 300 K using a weak restraint on the solute and a time step of 0.5 fs. A cut-off of 9 Å was used to compute the non-bonded interactions and Particle Mesh Ewald summation method (PME)[20] was used to deal with long-range. First the systems were heated at constant volume (NVT) at 150 K for 50 ps restraining the dimer positions with a k=20 kcal/molÅ$^2$. Then the solute restraint weights were set to 10 kcal/molÅ$^2$ and the two systems were equilibrated at 300 K in NVT condition for 50 ps followed by 50 ps at constant pressure (NPT, p= 1 bar). Finally, a last 10 ps equilibration NVT process was performed with no restrictions on the systems. The Berendsen's algorithm was used to control pressure with a relaxation time of 1.0 ps and the Langevin thermostat was employed with a collision frequency of 2 ps$^{-1}$. SHAKE[21] was used to constrain all the bonds involving hydrogen.

For the production step, five independent MD runs of 10 ns each were performed in NPT condition using a time step of 2 fs and the pmemd module of AMBER11. For each run temperatures were randomly chosen on the basis of a Maxwellian distribution at 300 K, while coordinates were taken for the first run from the structure of the equilibration step and for the following ones from the final structure of the previous 10 ns run. Structures for analysis were sampled every 10 ps and each 10 ns run concatenated resulting in a trajectory of 5,000 structures.

#### *6.6.3.3 MD results*

The trajectories obtained from the MD simulations were analyzed using the ptraj module of Amber11 package. To assess the stability of the dimers and the folding of each single domain, we analyzed the Root Mean Square Displacement (RMSD) of the backbone atoms Cα, C, N with respect to the input structure as a function of time. The EC1 (1-100 residues) and EC2 (101-218 residues) domains and the EC1-EC2 monomer forming the E- and N-cadherin dimers all showed little fluctuations of the backbone RMSD compared to the corresponding X-ray structure (RMSD < 2 Å for single EC1 or EC2 domains and RMSD< 3 Å for the 93% of simulation time for E-cadherin EC1-EC2 monomers and 99% for the N-cadherin EC1-EC2 monomers), i.e. the single domains seem to conserve the input folded structure and the monomer behaves like a rather rigid unit. Major RMSD fluctuations are observed for both E- and N-cadherin dimers, where the RMSD oscillated between 2 and 8 Å (Appendix B.1), showing a similar evolution of the corresponding dimer gyration radii (Appendix B.2). In fact, since compared to the X-ray structures we truncated our system to EC1-EC2 domains, some spatial rearrangements can occur. However these movements do not to interfere with the swap dimer interface interactions. In fact, EC1 centers of mass distance do not vary significantly during the simulation. The two partner molecules showed an average value of 22.5 Å and 23.4 Å for E- and N-cadherin, respectively, with 23.0 Å and 21.5 Å being the initial centers of mass distances in the input structure, for E- and N-cadherin, respectively.

## 6.7 BIBLIOGRAPHY

(1) Lipman, D.; Pearson, W. *Science* **1985**, *227*, 1435–1441.

(2) Smith, T. F.; Waterman, M. S. *J. Mol. Biol.* **1981**, *147*, 195–197.

(3) Harrison, O. J.; Jin, X.; Hong, S.; Bahna, F.; Ahlsen, G.; Brasch, J.; Wu, Y.; Vendome, J.; Felsovalyi, K.; Hampton, C. M.; Troyanovsky, R. B.; Ben-Shaul, A.; Frank, J.; Troyanovsky, S. M.; Shapiro, L.; Honig, B.; Sergey, M. *Structure* **2011**, *19*, 244–256.

(4) Shapiro, L.; Fannon, A. M.; Kwong, P. D.; Thompson, A.; Lehmann, M. S.; Grübel, G.; Legrand, J. F.; Als-Nielsen, J.; Colman, D. R.; Hendrickson, W. A. *Nature* **1995**, *374*, 327–337.

(5) Li, H.; Price, D. K.; Figg, W. D. *Anticancer. Drugs* **2007**, *18*, 563–568.

(6) Vendome, J.; Posy, S.; Jin, X.; Bahna, F.; Ahlsen, G.; Shapiro, L.; Honig, B. *Nat. Struct. Mol. Biol.* **2011**, *18*, 693–700.

(7) SiteMap, version 2.1, Schrödinger, LLC, New York, NY, **2007**.

(8) Laurie, A. T. R.; Jackson, R. M. *Bioinformatics* **2005**, *21*, 1908–1916.

(9) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849–857.

(10) Tuncbag, N.; Kar, G.; Keskin, O.; Gursoy, A.; Nussinov, R. *Brief. Bioinforma.* **2009**, *10*, 217–232.

(11) Kollman, P. a; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. a; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(12) Kortemme, T.; Kim, D.; Baker, D. *Sci. Signal.* **2004**, 1–8.

(13) Brasch, J.; Harrison, O. J.; Honig, B.; Shapiro, L. *Trends Cell Biol.* **2012**, *22*, 299–310.

(14) Case, D. A.; Darden, T. A. *AMBER 11*; 2010.

(15) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(16) Hendlich, M.; Rippmann, F.; Barnickel, G. *J. Mol. Graph. Model.* **1997**, *15*, 359–363.

(17) Maestro, version 9.2, Schrödinger, LLC, New York, NY, **2011**.

(18) http://ambermd.org/doc11/AmberTools.pdf.

(19) Bradbrook, G. M.; Gleichmann, T.; Harrop, S. J.; Habash, J.; Raftery, J.; Kalb (Gilboa), J.; Yariv, J.; Hillier, I. H.; Helliwell, J. R. *J. Chem. Soc. Faraday Trans.* **1998**, *94*, 1603–1611.

(20) Ewald, P. P. *Ann. Phys.* **1921**, *369*, 253–287.

(21) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. . *J. Comput. Phys.* **1977**, *23*, 327–341.

# Chapter 7: Virtual screening and design of cadherin inhibitors

## 7.1 INTRODUCTION

The analysis of the N- and E-cadherin dimer interfaces has brought to the conclusion that the most important features of the interaction are shared among these two types of cadherins. A few differences exist, however, in the residues forming the hydrophobic pocket of E-cadherin and N-cadherin, but they do not seem to affect the ability of type I classical cadherins to undergo heterophilic binding. Niessen and Gumbiner[1] have performed experiments on immobilized cadherin ectodomains and revealed substantial cross binding among N- and E-cadherin. Subsequent studies have confirmed this finding, while also pointing out that heterophilic binding affinities are intermediate between the homophilic affinities of the two classes of cadherins.[2,3]

So, in our effort to identify small molecules being able to inhibit both the N- and E-cadherin mediated cellular binding, we focused our attention on the interaction hot spots common to both cadherins. Finding small molecules able to disrupt the protein-protein interfaces is a challenging task.[4] Unlike the proteins having a natural small-molecule partner, proteins involved in PPIs reveal a much larger contact surface, that, in addition, is usually rather smoothed or flat. The computational means used to tackle this problem, for cadherin interactions, will be the subject of this chapter.

Essentially, our search for small cadherin inhibitors was achieved on one hand by performing a virtual screening of web server databases of commercially available compounds, in order to find mimics of the tetrapeptide adhesive arm, and on the other hand by de novo designing peptidomimetic molecules carrying all the necessary features identified by the in silico analysis discussed in Chapter 6. To assess the ability of all the compounds to fit into the Trp2 hydrophobic pocket, docking calculations in both E- and N-cadherin models were performed, using the new crystal structures 3q2v and 3q2w,[5] respectively. The set up and validation of the docking models are described in the Methods section. Docking results allowed the selection of the most promising candidates for the organic synthesis, in the case of the newly designed molecules, and for the purchase, in the case of the compounds identified by databases filtering. The results of biological assays performed on the synthesized molecules will also be presented.

## 7.2 VIRTUAL SCREENING OF DATABASES

### 7.2.1 PubChem

We started our investigation by first screening the PubChem[i] database. PubChem performs a 2D similarity search, measured using the Tanimoto equation[6] and the PubChem dictionary-based binary fingerprint.[7] Our search for compounds having at least 80 % similarity with respect to the tripeptide DWV produced ca. 40000 hits. The next step was to filter out these compounds based on a 3D pharmacophoric hypothesis derived from the characterization of the cadherin interface. We used Phase[8] and the DWV X-ray structure of N-cadherin (3q2w.pdb), which is very similar to that of the E-cadherin, to generate a five site 3D pharmacophoric query (Figure 56), that was exploited to filter the compounds after converting the 2D hits in three-dimensional structures.

[i] http://pubchem.ncbi.nlm.nih.gov/

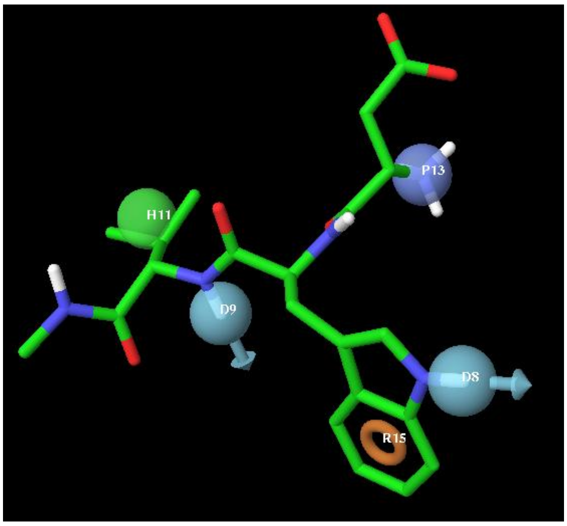


Figure 56. Five site 3D pharmacophoric hypothesis built on N-cadherin X-ray structure using Phase. The positive charge of N-terminus Asp1 (P13), the aromatic ring of Trp2 (R15), the H-bond donor of indole Trp2 (D8), the H-bond donor of backbone Val3 (D9) and the hydrophobic side chain of Val3 (H11) are shown.

We selected the best 200 compounds, ranked depending on their ability in satisfying the pharmacophoric requirements, measured using a fitness function.[9,10] Unfortunately, all the compounds had peptidic nature, with a Trp residue at position 2 (Appendix C.1) and thus do not provide a real alternative to the natural DWV sequence.

Docking calculations of these compounds into the N-cadherin model showed that some tripeptides are able to adopt the known 3D arrangement of the reference DWV sequence and form the key interactions within the binding site. This result reveals that the nature of the first and third amino acid seems not to affect the binding into the Trp2 pocket. However, due to the peptidic nature of the obtained ligands, we also performed the screening of a database of peptidomimetic molecules, which I will describe in the next section.

### 7.2.2 PEPMMsMimic

PepMMsMimic[j], developed by Moro and coworkers,[11] is a peptidomimetic database which comprises ca. 17 million 3D conformers, calculated from ca. 4 million unique compounds.[12] From the 3D structure of the tetrapeptide DWVI in the N-cadherin dimer, we generated a four site pharmacophoric model, by selecting the $NH_3^+$ group in Asp1, the side chain of Trp2, the NH group in the backbone of Val3 and the carbonyl group in the backbone of Asp1 (which correspond to interactions 1, 2, 4 and 5 described in Chapter 5). PepMMsMimic converts these selections into annotation points (the fingerprints) that define the 3D query. After the application of a search criterion (see below), the best 200 compounds are obtained in 2D format.

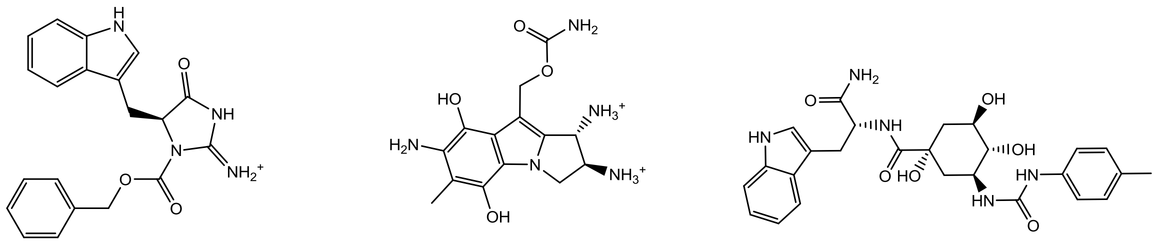
We separately used 3 criteria of search:

- a pharmacophoric fingerprint similarity
- a shape-based filtering of fingerprint similarity
- an hybrid method using both shape similarity and fingerprint similarity

At this point, 600 hits (200 x 3 data sets) were generated. However, a lot of duplicates were present among the three data sets. In addition, PepMMsMimic considers the conformers belonging to the same molecule as separate entries. A first filtering, based on the SMILES[13] sequence of each compound (Appendix C.2), was then performed to obtain ca. 200 unique compounds. We then used the LigPrep[14] module of the Schrödinger suite to convert the unique hits into 3D structure. Both E- and N-cadherin models were used for docking calculations, saving one pose per compound. Analysis of the molecules docked into the Trp2 pocket was performed independently for E- and N-cadherin. The poses were ranked on the basis of the Glide score. The compounds which did not insert their hydrophobic ring in the Trp2 pocket and did not form a salt bridge with the side chain carboxylate of Glu89 were filtered out. After docking filtering, the best compounds common to both the E- and N-cadherin and having non-

[j] http://mms.dsfarm.unipd.it/pepMMsMIMIC/

peptidic nature were selected as the best candidates, and are now being purchased for biological evaluation (Figure 57).



MMs02472555 MMs0344904013 MMs00030251

Figure 57. Best non-peptidic compounds extracted from PepMMsMimic as mimics of the tetrapeptide DWVI and top ranked in docking to both E- and N-cadherin.

## 7.3 Rational design of peptidomimetic cadherin inhibitors

Aiming at the modulation of the homophilic N- and E-cadherin binding with small peptidomimetic molecules, we designed a library of conformationally constrained mimics of the N-terminal tetrapeptide sequence Asp1-Trp2-Val3-Ile4 (DWVI) identified by the X-ray trans-swap dimer structures. The compounds of general formula $NH_3^+$-Asp-scaffold-Ile-$NHCH_3$ (see Appendix C.3 for the full list) were generated by replacing the central dipeptide Trp2-Val3 unit of the DWVI adhesive motif with several scaffolds developed in our laboratories.[15–18] The scaffolds are depicted in Figure 58. By changing the configuration of the relative stereogenic centers both the diketopiperazine (IV in Figure 58) and the bicyclolactame (I-III, V, VI in Figure 58) scaffolds allow variability in the spatial orientation of the substituents.

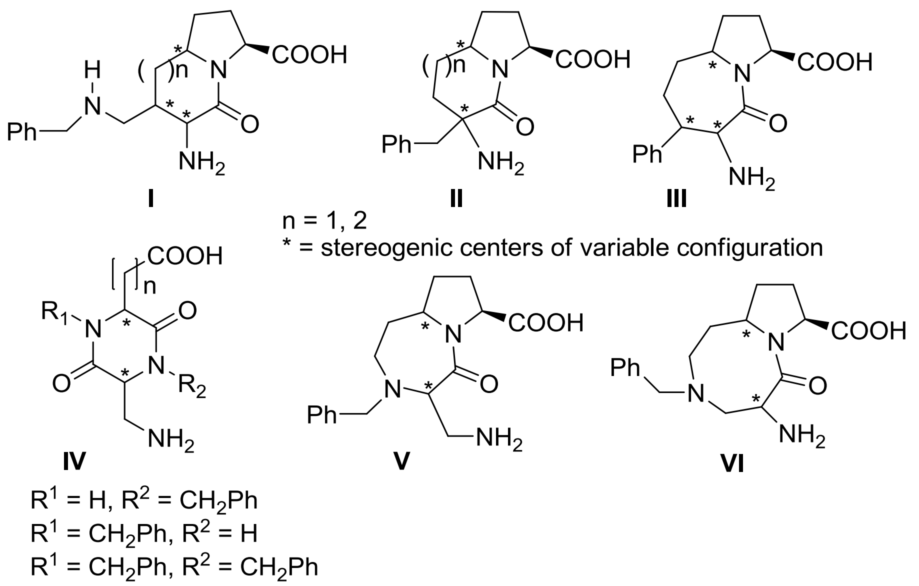


Figure 58. Scaffolds used for the generation of the virtual library of tetrapeptide mimics.

With respect to the natural ligand DWVI, the mimics, due to the cyclic nature of the scaffold, are conformationally constrained and due to their non-peptidic nature, more likely to survive metabolic degradation. This first generation library used a phenyl or benzyl moiety to mimic the indole moiety of Trp2 side chain, so no hydrogen bond can be formed within the binding site.

Compounds were docked into both E- and N-cadherin models, using Glide,[19] and saving ten poses for each ligand. The results were sorted on the basis of the Glide score and filtered to match the two most important binding interactions:

i. the formation of the salt bridge between the positively charged $NH_3^+$ group of the ligand and the side chain carboxylate of Glu89

ii. the anchoring of the hydrophobic ring of the ligand (either a phenyl or a benzyl group) into the hydrophobic pocket.

Differences in the binding poses of the compounds, when docked into N- and E-cadherin, only concern the positively charged $NH_3^+$ group, which is able to form an additional electrostatic interaction with the Asp27 side chain in the N-cadherin, while no additional salt bridge can be formed in the E-cadherin model (Asp27 is mutated to Asn27). No other significant difference was found in the 3D arrangement of the compounds when docked in the two different cadherin models.

Most of the compounds including the bicyclolactame scaffolds (Figure 58, type I-III, V) failed in reproducing the interaction (i) and (ii) or they matched the pose filtering criteria just in the top-ranked pose. Only peptidomimetic **1** (Figure 59) based on scaffold type VI of Figure 58, was able to form interaction (i) and (ii) for 3 poses over 10 in the E-cadherin and for 4 over 10 in the N-cadherin (Table 8).

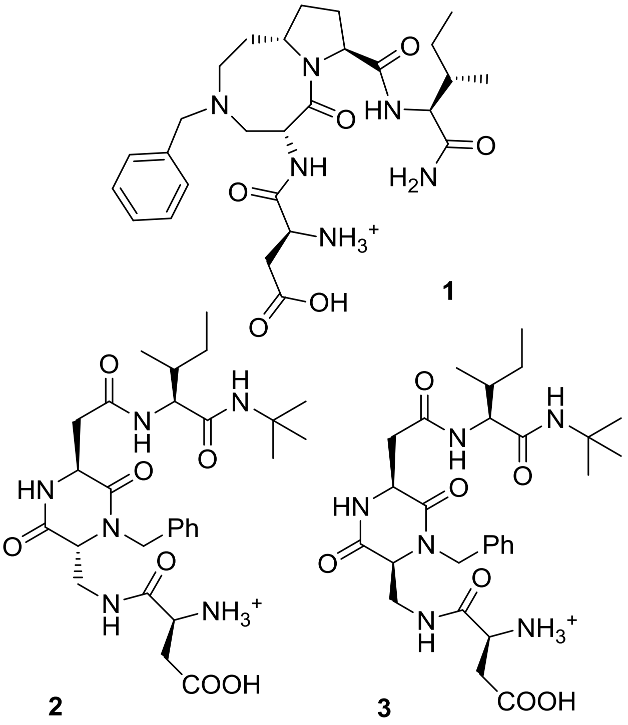


Figure 59. Peptidomimetics ranked best among the designed ligands.

| E-cadherin | | | |
|---|---|---|---|
| **Compound** | **Glide score range** | **Asp1-NH$_3^+$/C$^d$OO$^-$Glu89*** | **benzyl ring in hydrophobic pocket** |
| **DWVI** | -11.64 / -9.58 | 10/10 | 10/10 |
| **1** | -7.06 / -5.94 | 3/10 | 3/10 |
| **2** | -7.34 / -6.10 | 7/10 | 10/10 |
| **3** | -9.10 / -6.93 | 9/10 | 10/10 |

*distance between N and C < 4.0 Å

| N-cadherin | | | |
|---|---|---|---|
| **Compound** | **Glide score range** | **Asp1-NH$_3^+$/C$^d$OO$^-$Glu89*** | **benzyl ring in hydrophobic pocket** |
| **DWVI** | -10.52 / -7.34 | 10/10 | 10/10 |
| **1** | -6.98 / -5.47 | 4/10 | 4/10 |
| **2** | -8.74 / -6.73 | 5/10 | 10/10 |
| **3** | -7.99 / -5.19 | 7/10 | 10/10 |

Table 8. Docking results for the compounds **1**, **2** and **3**. 10 poses were saved for each compound. Results for the tetrapeptide DWVI are reported as reference. Ligands were docked to both E-cadherin (top) and N-cadherin (bottom).

However, as can be seen in Figure 60, the binding mode of **1** in both receptors showed a different disposition of the ligand compared to the DWVI sequence. In particular, **1** orients the Ile residue back to the Asp1 amino acid and does not reproduce the experimental backbone arrangement.

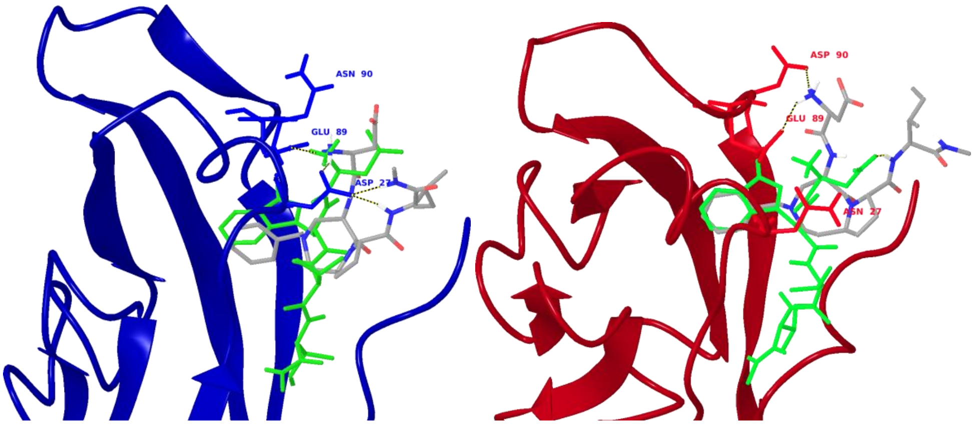


Figure 60. Best pose of **1** (tube representation, C in grey, N in blue and O in red) into the N- (left) and E-cadherin (right) models, overlaid to the DWVI sequence (green tube representation). Key receptor residues are labeled and highlighted in tube representation.

On the contrary, peptidomimetics containing the diketopiperazine scaffolds (Figure 58, type IV) showed generally better results according to both the Glide score and the number of poses reproducing the two interactions**Errore. L'origine riferimento non è stata trovata.**. Among them, compounds **2** and **3** (Figure 59) were able to form the interaction (i) for at least 5 over 10 poses, and the interaction (ii) for all the poses in both E- and N-cadherin models (Table 8). With respect to the reference tetrapeptide sequence DWVI, **2** and **3** were also able to overlay to the backbone X-ray structure (Figure 61).

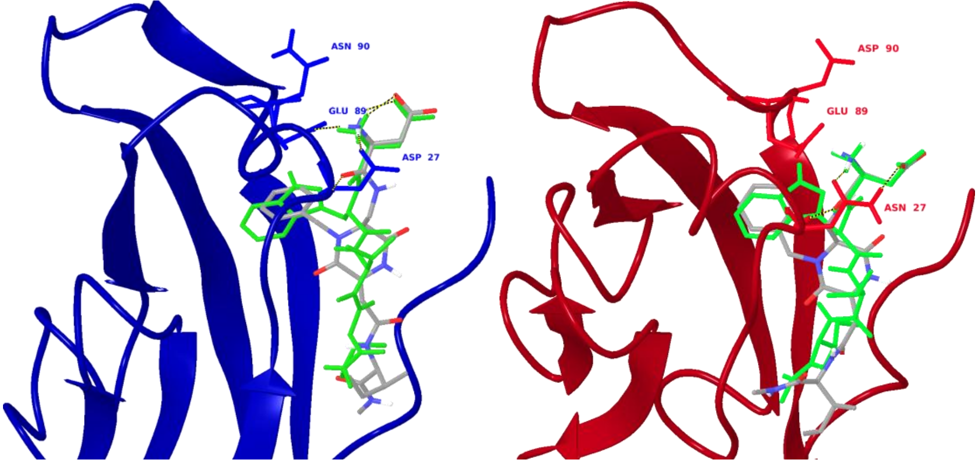


Figure 61. Best pose of **3** (tube representation, C in grey, N in blue and O in red) into the N- (left) and E-cadherin (right) models, overlaid to the DWVI sequence (green tube representation). Key receptor residues are labeled and highlighted in tube representation.

According to this analysis, the best compound obtained using a bicyclolactame scaffold (**1** in Figure 59), and the two most promising compounds having a diketopiperazine scaffold (**2** and **3** in Figure 59) were selected for the organic synthesis.

## 7.4 Biological assays and comparison with the in silico predictions

Compounds **1**, **2** and **3** were tested at the Istituto Nazionale Tumori by Dr. Antonella Tomassetti in cell adhesion assays using Epithelial Ovarian Cancer (EOC) cell lines expressing E- (OAW42) and N-cadherin (SKOV3). In the adhesion assays, cells were allowed to form monolayer in presence of each ligand at 2 different concentrations (2 and 1 mM). Comparative experiments were performed using ADH-1, the reference peptide already used in Phase I clinical trial for the treatment of ovarian cancer (see Chapter 5), and control experiments in the absence of any ligand. All compounds inhibited the formation of cell monolayers of N-cadherin-expressing cells at 2 mM concentration, comparably to ADH-1. Noteworthy, **2** and **3** were also active at 1 mM concentration, and **3** was able to inhibit cell–cell aggregation of cells in suspension (Figure 62, left) On the other hand, all compounds, when tested on the E-cadherin-expressing cell line, showed lesser efficiency in inhibiting the formation of cell monolayer (Figure 62, right).

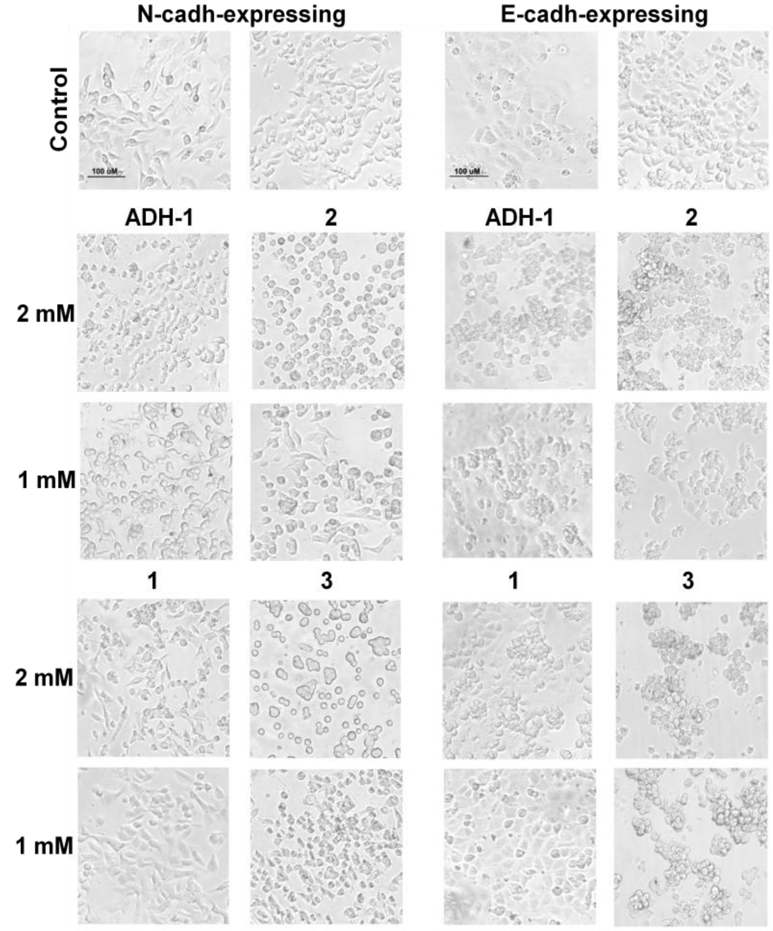


Figure 62. Adhesion assay to evaluate the inhibition of the formation of the cell monolayer by the small peptidomimetic ligands. N-cadherin (SKOV3) or E-cadherin (OAW42) expressing cells were seeded in absence (Control) or in presence of the ligands at 2 and 1 mM.

The compounds were also tested by enzyme-linked immunosorbent assay (ELISA)[20,21] for their ability to inhibit calcium-dependent cadherin homophilic interactions using the N- and E-cadherin-expressing cells and N- or E-cadherin-Fc chimeric protein, respectively (Figure 63). **2** and **3** ligands at 2 mM concentration inhibited N-cadherin homophilic binding by 78% and 84%, respectively, and 50% and 65% at 1 mM concentration (Figure 63, left). ADH-1 and **1** showed about 50% inhibition of N-cadherin/N-cadherin interactions at the higher concentration, and appeared ineffective at the lower concentration (Figure 63, left). Again, in the same test performed on the E-cadherin-expressing EOC cell lines, all compounds showed lesser efficiency in inhibiting E-cadherin homophilic interactions, with relevant effects only at the higher concentration (Figure 63, right).

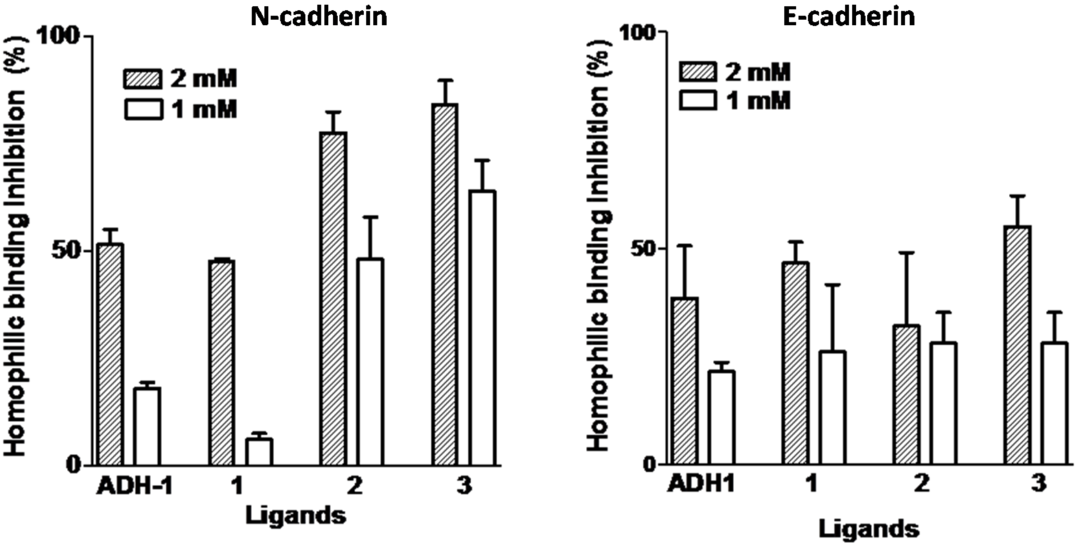


Figure 63. Left: inhibition of N-cadherin homophilic interaction by the small peptidomimetic ligands. Right: inhibition of E-cadherin homophilic binding by the small peptidomimetic ligands. The inhibition by ADH-1 is reported as control. The graphs report the mean values ±SD.

In fact, by ELISA 2 mM **1** and **3** inhibited E-cadherin/E-cadherin interaction by about 50% while ADH-1 showed only 30% inhibition (Figure 63, right), indicating a slight better efficacy compared to ADH-1 in inhibiting also E-cadherin homophilic interaction.

Combining the computational investigations with the results of the cell adhesion assays may provide a significant contribution to our understanding of the ligand structural requirement for the inhibition of N-cadherin homophilic interactions. Since only **2** and **3** were able to significantly inhibit N-cadherin-mediated adhesion in EOC cells (SKOV3), it appears that small molecules modulators of N-cadherin homophilic binding can reproduce the key interactions (i) and (ii) and also align to the DWVI backbone.

Although N- and E-cadherins share similar adhesive binding features for the DWVI sequence,[5] and marked differences in the interaction mode of our peptidomimetic ligands were not observed in the two cadherin docking models, the tests on the E-cadherin-expressing EOC cell line OAW42 showed lower inhibition capability of E-cadherin homophilic interactions compared with those of N-cadherin. However, two-dimensional affinities measured by micropipette adhesion assays on cell lines expressing E- or N-cadherins,[22] have shown that E-cadherin homophilic binding is stronger than that of N-cadherin. For this reason, our compounds might be less efficient in inhibiting E-cadherin than N-cadherin interaction. Furthermore, according to several compelling data, we focused our investigation exclusively on the inhibition of the trans-swap dimer formation targeting the EC1 domain, although several extracellular domains are known to be involved in the adhesive interface.[22] In addition, cadherin cis interactions contributing to the strength of cell-cell adhesion, could be another factor negatively affecting the efficacy of our ligands.

## 7.5 Conclusions

Targeting the interfaces between proteins has huge therapeutic potential, and discovering small drug-like molecules able to modulate protein-protein interactions is of major interest, but a difficult challenge of research. In this chapter I have presented the first attempt to rationally design small molecules targeting the trans-swap dimer interfaces of N- and E-cadherin. Remarkably, two of our peptidomimetics have shown to inhibit the N-cadherin-mediated adhesion process in EOC cells with improved efficacy in comparison with the ADH-1 cyclic peptide, already investigated as an N-cadherin antagonist in phase I clinical studies in various tumors, [23] including EOCs.[24]

In the next future we intend to improve the affinity of our ligands for either E- or N-cadherin, or both, by optimization of the interactions with the receptor. A possible modification could be located at the aromatic ring, by introducing the appropriate functionalities in order to promote the formation of a hydrogen bond within the Trp2 pocket. Other hints could be provided by the compounds selected by the database screening.

In addition, NMR and crystallographic studies will be performed to improve the structural understanding of the binding of our peptidomimetic ligands to the extracellular domains of the cadherins.

Thus, these small molecules represent a lead for a new class of modulators of cadherin-mediated adhesion, as important tools in the investigation of cellular processes, and in the design of novel diagnostic and therapeutic approaches for tumors, especially for EOCs.

## 7.6 METHODS

### 7.6.1 Set up and validation of the docking models

The automated docking calculations were performed using Glide.[19] Proteins (3q2w and 3q2v for N- and E-cadherin, respectively) were prepared as described in the Methods section of Chapter 6, using the Protein Preparation Wizard of the Maestro Graphical User Interface.[25] Models were set up by selecting only the EC1 domain (residues 1-103 and two $Ca^{2+}$ ions at the end of EC1) as receptor and the DWVIPP esapeptide sequence as ligand. The center of the grid enclosing box was defined by the center of the DWVIPP sequence. The enclosing box dimensions, which are automatically deduced from the ligand size, fit the entire active site. Docking calculations were performed using the standard precision mode (SP). The receptor was considered as a rigid body while the ligand sampling was set to 'Flexible' with the option 'Penalize non planar conformation' for amides. No Epik state penalties were used in the docking score calculations. The size of the bounding box for placing the ligand center was set to 14 Å. No further modifications were applied to the default settings. The Glide score[26] function was used to select 10 poses for each ligand.

Validation of the docking models was performed by testing the ability of reproducing the crystallized binding mode of fragments of the N-terminal native sequence, from the tripeptide DWV up to the decapeptide (DWVIPPINLP and DWVIPPISCP for N- and E-cadherin, respectively). The program was successful in reproducing the experimentally determined binding mode of these peptides. Key crystallographic contacts of the homophilic interaction (described in Chapter 6) were reproduced in the top poses of each peptide. In Table 9 the docking results for the tetrapeptide DWVI are summarized. In this case, all poses reproduce the most important interactions, both in the E-cadherin and the N-cadherin.

| Interaction (DWVI-Receptor Residues) | Number of poses | |
|---|---|---|
| | N-cadherin | E-cadherin |
| Asp1-$NH_3^+$/$C^{\delta}OO^-$Glu89* | 10/10 | 10/10 |
| Trp2 side chain in pocket | 10/10 | 10/10 |
| Trp2-$N^{\varepsilon 1}$H--CO-Asn90$_{N\text{-}cadh}$/CO-Asp90$_{E\text{-}cadh}$** | 10/10 | 10/10 |
| Val$^3$NH--COArg$_{N\text{-}cadh}^{25}$/COLys$_{E\text{-}cadh}^{25}$** | 10/10 | 9/10 |
| Asp1-CO--NH-Asp27$_{N\text{-}cadh}$/NH-Asn27$_{E\text{-}cadh}$** | 9/10 | 10/10 |

*distance between N and C < 4.0 Å,** distance between H and O < 2.5 Å

Table 9. Redocking of DWVI onto N- and E-cadherin. All important interactions are maintained. Glide score for DWVI in N-cadherin ranges from -10.52 to -7.34. Glide score for DWVI in E-cadherin ranges from -11.64 to 9.58.

In Appendix C.4, a picture of the tripeptide best pose docked onto the N-cadherin is reported.

### 7.6.2 Virtual screening of databases

3D pharmacophores, used both for Phase and PepMMsMimic, were generated based only on the 3D structure of the DWVIPP peptide in the 3q2w N-cadherin dimer structure. The difference in RMSD between the tetrapeptide in the N-cadherin dimer and the same peptide in the E-cadherin dimer ranges from 0.73 and 0.75 Å (the two monomer in E-cadherin, contrary to N-cadherin, are not symmetric), so that no important difference in the relative pharmacophoric site distances arises.

2D structures from the virtual databases were converted to 3D using LigPrep.[14] For each 2D structure, only one 3D conformer was generated, without changing its ionization state. Each compound was minimized with a OPLS_2005 force field.[27]

### 7.6.3 Virtual screening of tetrapeptide mimics

The library of DWVI peptidomimetics was evaluated in the E-cadherin and N-cadherin models using the same protocol of the validation step. 10 poses for each compounds were saved and analyzed considering the Glide docking score and the x-ray reference interaction models of the DWVI sequence. In particular, we filtered the generated poses using the (i) and (ii) interactions criteria.

## 7.7 BIBLIOGRAPHY

(1) Niessen, C. M.; Gumbiner, B. M. *J. Cell Biol.* **2002**, *156*, 389–399.

(2) Prakasam, A. K.; Maruthamuthu, V.; Leckband, D. E. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15434–154349.

(3) Niessen, C. M.; Leckband, D.; Yap, A. S. *Physiol. Rev.* **2011**, *91*, 691–731.

(4) Wells, J.; McClendon, C. L. *Nature* **2007**, *450*, 1001–1009.

(5) Harrison, O. J.; Jin, X.; Hong, S.; Bahna, F.; Ahlsen, G.; Brasch, J.; Wu, Y.; Vendome, J.; Felsovalyi, K.; Hampton, C. M.; Troyanovsky, R. B.; Ben-Shaul, A.; Frank, J.; Troyanovsky, S. M.; Shapiro, L.; Honig, B.; Sergey, M. *Structure* **2011**, *19*, 244–256.

(6) Rogers, D. J.; Tanimoto, T. T. *Science* **1960**, *132*, 1115–1118.

(7) ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf.

(8) Phase, version 3.3, Schrödinger, LLC, New York, NY, **2011**.

(9) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. *Chem. Biol. Drug Des.* **2006**, *67*, 370–372.

(10) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. *J. Comput. Aided. Mol. Des.* **2006**, *20*, 647–671.

(11) Floris, M.; Masciocchi, J.; Fanton, M.; Moro, S. *Nucleic Acids Res.* **2011**, *39*, W261–269.

(12) Masciocchi, J.; Frau, G.; Fanton, M.; Sturlese, M.; Floris, M.; Pireddu, L.; Palla, P.; Cedrati, F.; Rodriguez-Tomé, P.; Moro, S. *Nucleic Acids Res.* **2009**, *37*, D284–290.

(13) Weininger, D. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(14) LigPrep, version 2.5, Schrödinger, LLC, New York, NY, **2011**.

(15) Manzoni, L.; Arosio, D.; Belvisi, L.; Bracci, A.; Colombo, M.; Invernizzi, D.; Scolastico, C. *J. Org. Chem.* **2005**, *70*, 4124–4132.

(16) Ressurreição, A. S. M.; Bordessa, A.; Civera, M.; Belvisi, L.; Gennari, C.; Piarulli, U. *J. Org. Chem.* **2008**, *73*, 652–660.

(17) Arosio, D.; Belvisi, L.; Colombo, L.; Colombo, M.; Invernizzi, D.; Manzoni, L.; Potenza, D.; Serra, M.; Castorina, M.; Pisano, C.; Scolastico, C. *ChemMedChem* **2008**, *3*, 1589–1603.

(18) Manzoni, L.; Belvisi, L.; Arosio, D.; Civera, M.; Pilkington-Miksa, M.; Potenza, D.; Caprini, A.; Araldi, E. M. V; Monferini, E.; Mancino, M.; Podestà, F.; Scolastico, C. *ChemMedChem* **2009**, *4*, 615–632.

(19) Glide, version 5.7, Schrödinger, LLC, New York, NY, **2011**.

(20) Engvall, E.; Perlmann, P. Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry* **1971**, *8*, 871–874.

(21) Van Weemen, B. K.; Schuurs, A. H. W. M. *FEBS Lett.* **1971**, *15*, 232–236.

(22) Leckband, D.; Sivasankar, S. *Curr. Opin. Cell Biol.* **2012**, *24*, 620–627.

(23) Blaschuk, O. W. *Cell Tissue Res.* **2012**, *348*, 309–313.

(24) Perotti, A.; Sessa, C.; Mancuso, A.; Noberasco, C.; Cresta, S.; Locatelli, A.; Carcangiu, M. L.; Passera, K.; Braghetti, A.; Scaramuzza, D.; Zanaboni, F.; Fasolo, A.; Capri, G.; Miani, M.; Peters, W. P.; Gianni, L. *Ann. Oncol.* **2009**, *20*, 741–745.

(25) Maestro, version 9.2, Schrödinger, LLC, New York, NY, **2011**.

(26) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput. Aided. Mol. Des.* **1997**, *11*, 425–445.

(27) MacroModel, version 9.5, Schrödinger, LLC, New York, NY, **2007**.

# Chapter 8: Study of the binding mechanism of E-cadherin

## 8.1 INDUCED FIT AND SELECTED FIT MECHANISMS

In Chapter 5, it was mentioned that classical cadherins dimerize through the mutual exchange of the adhesion arm, and the Trp2 side chain is inserted onto an acceptor pocket of the opposed cadherin coming from a different cell.

This binding mode is a typical example of the so-called three-dimensional domain swapping, introduced by Bennett and coworkers in 1995,[1] in which a domain of a protein, containing an hinge loop, is exchanged by an identical domain coming from another protein, forming a dimer. In many cases, the domain could simply be a single secondary structure element, like a β-strand or an α-helix. A common structural feature of domain swapping is the presence of prolines in the hinge loops. It has been speculated that the presence of these prolines causes strain in the loop and the release of this strain, leading to the swapping event, could be one of the driving forces behind 3D domain-swapping.[2]

Vendome and coworkers[3] proposed in 2011 that the driving force behind the arm opening of E-cadherin was the release of the strain caused by the anchoring of the arm at two far distant points, namely the already mentioned hydrophobic pocket, to which Trp2 side chain is bound in the closed form, and the $Ca^{2+}$ ions at the EC1-EC2 boundary, with which many negatively charged EC1 amino acids form electrostatic interactions. Thus, the release of the indole moiety from its pocket should be entropically driven (Figure 64). However, several aspects could affect this arm opening, most of them still requiring to be investigated. For instance, although E-cadherin adhesive arm contains two prolines, the mutation into alanine increased the dimerization affinity by almost two orders of magnitude and led to the formation of a novel type of swap dimer interface. Moreover, $Ca^{2+}$ ions at the interface of EC1-EC2 domains seem to activate the dimerization process.

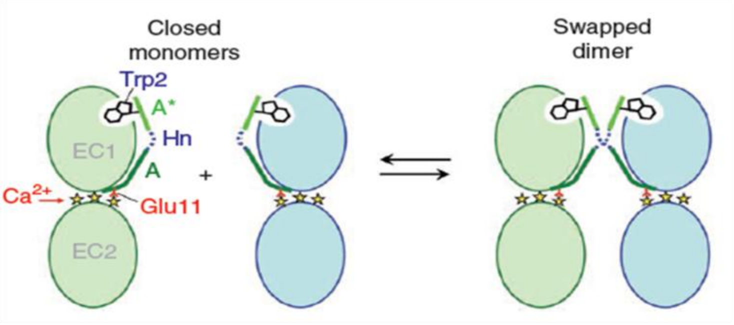


Figure 64. Left: anchor points (hydrophobic pocket and $Ca^{2+}$ ions) in closed form cadherins. Glu11 side chain, at the end of the trans-swap domain, forms a salt bridge with $Ca^{2+}$. Right: fully swapped dimer.

As noted earlier (Chapter 5), there is still debate regarding the path through which the swap dimer is formed starting from a cadherin monomer in the closed, inactive form. Two mechanisms have been proposed on the basis of experimental data (Figure 65).

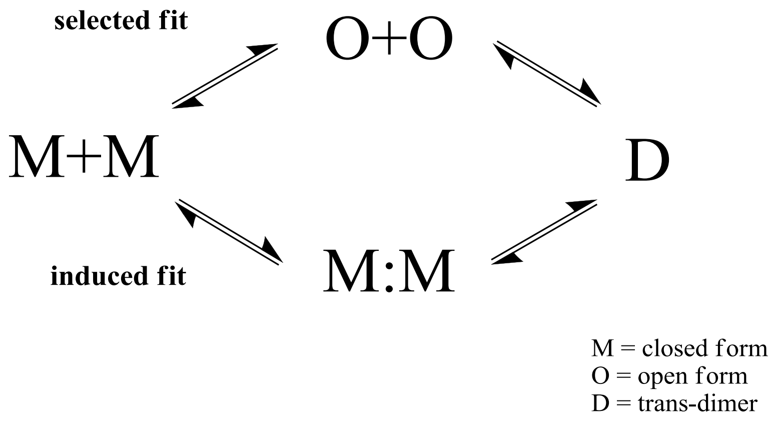


Figure 65. Proposed mechanisms for the dimerization of classical cadherins, selected fit (up) and induced fit (down) mechanisms. M:M stands for the encounter complex, the X-dimer.

Very few computational studies regarding cadherin binding mechanism have been published in the recent years[3–5]. In fact, the timescale involved in the formation of the dimer (in the millisecond range), limits any attempts at simulating cadherin conformational changes with conventional MD classical approaches. To overcome this problem we made use of an enhanced sampling technique, metadynamics (Chapter 2), that has proven itself to be able of studying a similarly complex molecular association process at a fully atomistic level.[6]

We used a combination of the parallel tempering and well-tempered metadynamics methods (PTMetaD), applying the well-tempered ensemble (WTE) methodology, to test the selected fit mechanism in E-cadherin.

## 8.2 METADYNAMICS SIMULATIONS

In this chapter the results obtained from metadynamics simulations of the E-cadherin conformational change will be presented.[k] These studies have been performed during my stay at the Centro Nacional de Investigaciones Oncológicas (CNIO), in Madrid, under the supervision of Prof. Francesco Luigi Gervasio.[l]

We simulated the first step of the selected fit mechanism, the opening of the arm (Figure 66), leading to a free open form of the E-cadherin monomer in solution.

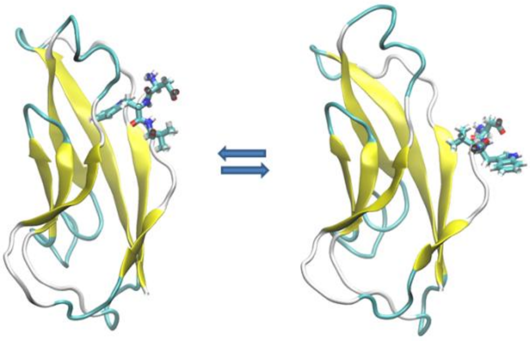


Figure 66. First step of the selected fit mechanism, as depicted in Figure 65. Here, only EC1 E-cadherin is shown.

[k] Reconstructing the free energy landscape of E-cadherin conformational transition by atomistic simulations.; Doro, F.; Saladino, G.; Belvisi, L.; Civera, M.; Gervasio, F. L.; *manuscript in preparation*

[l] Present address: University College London, Department of Chemistry and Institute of Structural and Molecular Biology, London (United Kingdom)

We also examined the impact of $Ca^{2+}$ ions in favoring the conformational transition. To do that, we set up two different systems:

1. EC1 E-cadherin monomer in the closed form, with no $Ca^{2+}$ ions
2. EC1-EC2 E-cadherin monomer in the closed form containing three $Ca^{2+}$ ions at the EC boundary and a fourth calcium ion at the end of EC2.

Calcium ions in cadherins protect from proteolysis and rigidify the ectodomain but, as mentioned before, they could also be involved in the arm opening process.

E-cadherin was used for these mechanistic studies, rather than N-cadherin, since it is regarded as the prototypical type I classical cadherin. Moreover, a 3D structure of a closed monomer form only exists for E-cadherin.[7]

### 8.2.1 Preliminary metadynamics runs: choice of the best set of CVs

In metadynamics, convergence of the simulation is often dependent upon the Collective Variables (CVs) being chosen to describe the system. As a consequence an essential component in every MetaD simulation is performing a series of preliminary MetaD runs in order to find the optimal set of CVs. A series of 20 ns short MetaD simulations were performed for this purpose. E-cadherin conformational change in such a small time could be achieved by adding to the total potential, at every MetaD step, gaussians with the considerable height of 5 kcal/mol. In this way, we could favor the conformational change and select those CVs that better discriminate the closed form of E-cadherin from its open form.

By noting that the distance between the Trp2 side chain and the residues of the binding pocket (residues 73-80) greatly varies between the two forms, a first CV (CV1) was set to be the distance between the His79 Cα and the centroid of the indole moiety in Trp2 (Figure 67).

Then, it was needed a quantity that could describe the relative orientation of the Trp2 indole moiety with respect to the backbone of the protein. After careful evaluation, we opted to use as CV the dihedral between the principal inertia axis of the protein and the Trp2 indole ring (CV2, Figure 68), identified by two heavy atoms in the Trp indole ring and the centroids of two sets of Cα atoms in the β-strand regions 73-80 and 92-97.

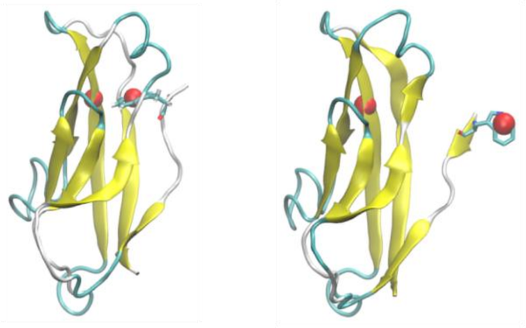


Figure 67. **CV1**: distance between His79 Cα and Trp2 indole centroid. In the closed conformation (left) CV1 is ca. 5.5 Å. In the open conformation CV1 ranges from 10 to 20 Å.

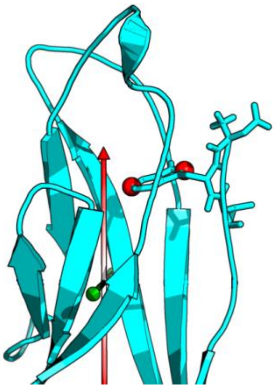


Figure 68. **CV2**: orientation of Trp2 indole ring with respect to the principal inertia axis of the protein (in red).

At this stage it was also attempted to use as third Collective Variable the so-called Coordination Number, in which the number of water molecules surrounding the Trp2 indole ring, in the first shell of solvation, was counted. Although the CV could significantly distinguish specific conformations of the E-cadherin during the closed-open conformational change (when inside the hydrophobic pocket, Trp2 indole is hindered from the solvent), the computational effort required to count at every MetaD step the number of molecules around the indole ring, excessively slowed the simulation, and an alternative choice had to be made. In Appendix D.1 the reconstructed free energy using this specific CV is included.

As a consequence, the third Collective Variable, introduced to better visualize the opening mechanism, was selected on the basis of the analysis of snapshots of closed, intermediate and open forms of E-cadherin taken from these short MetaD simulations. Careful analysis of the different atom distances between the hydrophobic pocket and the opening arm, in the three forms, allowed to define a Contact Map (CV3), describing the path along which the conformational change occurs. The contact map is defined as the sum of the contacts between a number of atom pairs, and each contact is defined using a switching function:[8]

$$D_{ab}(X) = \theta(c_{ab} - r_{ab})w_{ab}\frac{1-(r_{ab}/r_{ab}^{(0)})_{ab}^{n}}{1-(r_{ab}/r_{ab}^{(0)})_{ab}^{m}} \qquad (34)$$

where $a$ and $b$ refer to two set of atoms, $\theta$ is a step function, and $c_{ab}$, $r_{ab}^{(0)}$, $n_{ab}$ and $m_{ab}$ are parameters defining the reference distance for each atom set.

We chose a set of contacts so that the entire opening process could be observed. In Table 10 the contacts being used, together with their reference value, are shown. So, a contact map value of 3 describes a closed conformation, while a contact map value of 1 is obtained for an open conformation.

| contact | conformation | reference value (Å) |
|---|---|---|
| Asp1 - Glu89 | closed | 3.7 |
| Trp2 - Met92 | closed | 4.1 |
| Trp2 - Asp90 | closed | 2.0 |
| Trp2 - Glu89 | intermediate | 3.2 |
| Asp1 - Asp90 | intermediate | 5.9 |
| Trp2 - His79 | open | 20.0 |

Table 10. **CV3**: Contact Map. Contacts belonging to the closed conformation are labeled as *closed*. Two contacts appearing during the opening of the arm are labeled as *intermediate*. One last contact describes the *open* conformation.

### 8.2.3 Parallel Tempering Metadynamics in the Well-Tempered Ensemble

In WTE-PTMetaD, one first has to prepare a small number of replicas of the system. As an initial step, a Parallel Tempering metadynamics is performed using only the potential energy as CV. This increases the overlap in the potential energy distribution between replicas at adjacent temperatures, thus permitting to reach a high exchange probability even with the use of only a small number of replicas. In our case 4 replicas were generated, spanning a temperature range of 300-400 K, and PT-MetaD runs enabled an exchange probability of ca. 30 % for both EC1 and EC1-EC2 systems.

After this step, the obtained metadynamics bias potential is implemented as a fixed potential in subsequent production runs of WTE-PTMeatD, with the effect of greatly amplifying the fluctuations in the potential energy. The simulation temperatures for replicas were set to 300 K, 330 K, 362 K and 398 K, running for 200 ns (for EC1) and 250 ns (for EC1-EC2). GROMACS v 4.5.5 with PLUMED v 1.3 plugin was used. Simulation details are described in the Methods section.

## 8.3 RESULTS

### 8.3.1 Convergence of WTE-PTMetaD calculation

Following Deighan and coworkers,[9] we assessed the convergence of our simulations by observing three important facts:

1. the interesting regions in the CV space were fully explored
2. the trajectories along the CV dimensions showed diffusion through the closed and open forms, so the system crosses the energy barrier and the conformational transition occurs reversibly
3. the free energy differences between the two main minima (closed and open form) remained constant.

Simulations were stopped when all the above three requirements were satisfied. In Figure 69 the free energy difference of two main minima, for EC1 and EC1-EC2 systems, is shown, while in Appendix D.2 the script used to check this convergence is reported.

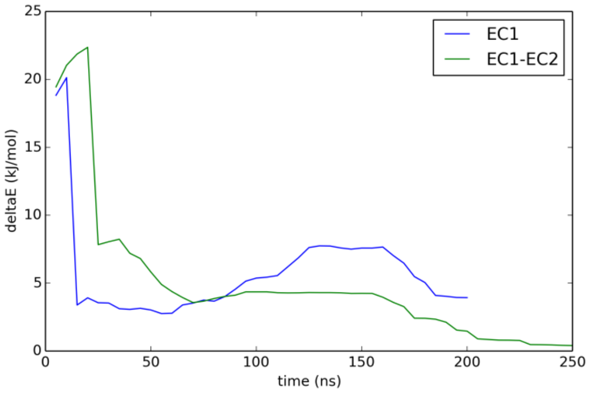


Figure 69. Time series of the free energy differences between closed state and open state in both EC1 and EC1-EC2 cadherin. Data points are collected every 5 ns.

### 8.3.2 Outline of EC1 and EC1-EC2 Free Energy profiles

The calculated free energies at 300 K, obtained from the history-dependent biasing potential of WTE-PTMetaD and projected on a 2D surface using CV1 and CV2, are shown in Figure 70 and Figure 71. In both pictures, a minimum (minimum A in the pictures) is located at values [6.5 Å, π/2 rad], representing the closed conformation, with Trp2 side chain being firmly docked onto its hydrophobic pocket and with very limited conformational flexibility. Upon the opening of the adhesive arm, the Trp side chain is relatively free to move in solution, and can adopt many different conformations. However, since minima along the CV2 variable differ only by the orientation of the indole ring with respect to the protein principal axis, minima can be grouped on the basis of their CV1 value.

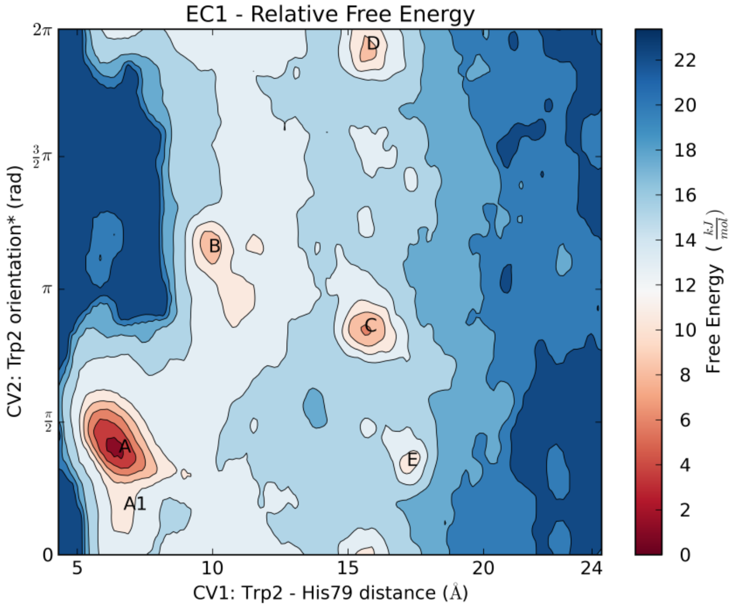


Figure 70. EC1 - Reconstructed Free Energy projected on a 2D surface using CV1 and CV2.

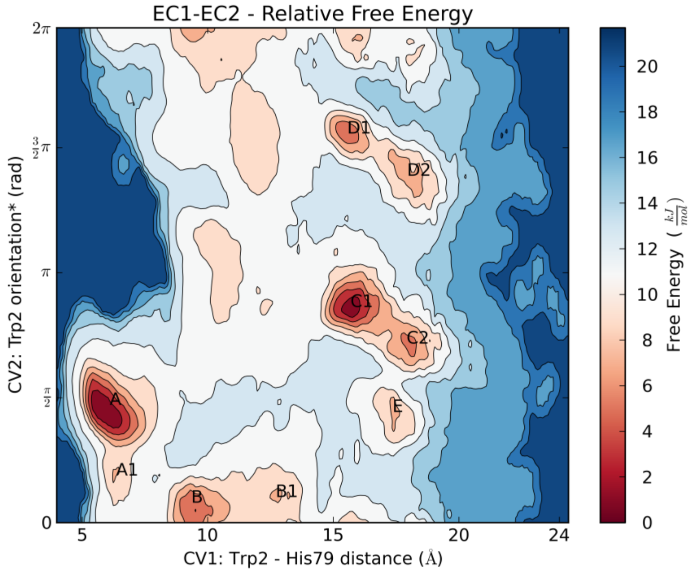


Figure 71. EC1-EC2 - Reconstructed Free Energy projected on a 2D surface using CV1 and CV2.

For instance, the minima C and D of Figure 70 differ only for the orientation of the Trp2 indole moiety. For this reason, free energy profiles projected solely on CV1 permit an easier comparison. Analyses of both EC1 and EC1-EC2 1D free energy profiles (Figure 72) show the presence of two major open form minima besides the closed form.

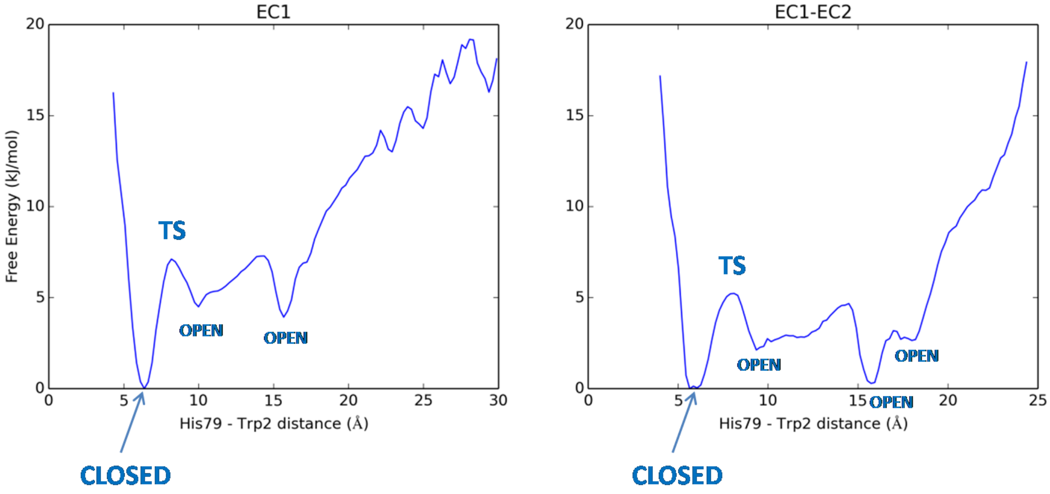


Figure 72. Free energy profiles projected on CV1 for both EC1 (left) and EC1-EC2 (right).

In the first open form located at ca. 10 Å in the CV1 space (minimum B in Figure 70 and in Figure 71), the Trp2 side chain tends to adhere to the rest of the protein in order to minimize the solvent exposure and does not resemble the conformation adopted in the trans-swapped dimer. The second open minimum at ca. 16 Å (minimum C in Figure 70 and minimum C1 in Figure 71) is fully open and the Trp2 side chain is completely exposed to the solvent. Only for the EC1-EC2 system we observe a third open minimum located at ca 18 Å (corresponding to the minimum C2 in Figure 71) that effectively represents the E-cadherin open form as seen in the swapped-dimer crystal structure of E-cadherin (Figure 73).

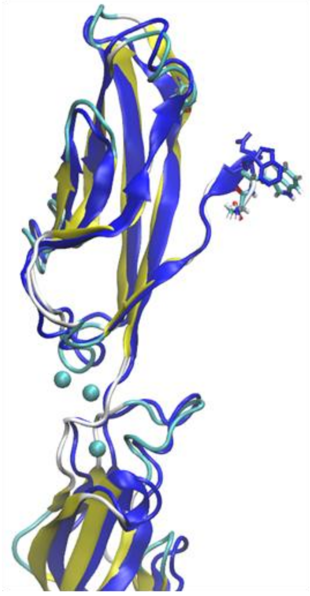


Figure 73. Superposition between EC1-EC2 E-cadherin (pdb code 3q2v) and a representative conformation of C2 minimum from Figure 71. RMSD (computed on the heavy atoms of residues 1-6) is 0.9 Å.

From the energetic point of view, in the EC1 system the two open form minima are at an unfavorable energy with respect to the corresponding closed form minimum. This result is somewhat expected as Miloushev and coworkers have experimentally measured the ΔG for a similar equilibrium and found that the closed form of type II cadherin 8 is ca. 2 kcal/mol more stable than the open form.[10] In our case, differences in the minima amount to ca. 1 kcal/mol, probably due to the different nature of the strand-swapping adhesive arm (in type II cadherin 8, two stacked indole moieties are exchanged during strand swapping). For the EC1-EC2 system, the two open minima are at a more favorable energy, and the main open form conformation is isoenergetic with respect to the closed form. Moreover, we also observed a lowering of the transition state energy (by 0.5 kcal/mol, as depicted in Figure 72) and, as a consequence, of the energy barrier of the closed-to-open transition.

In general, the opening of the arm should be an entropy-driven process, since the arm leaves the binding pocket, where it adopts a well-defined conformation, becoming exposed to the solvent and then conformationally free. Looking at the CV2 space (that samples the conformational flexibility of the Trp2 indole moiety), for the EC1 system we obtained the unusual result of locating only few open forms (Figure 70, minima B-E). Moreover, none of such open forms correspond to the structure found in E-cadherin swap-dimer. On the contrary, the EC1-EC2 system results in a broader selection of open conformations and, only for this system containing the $Ca^{2+}$ ions, an open form corresponding to the X-ray swap dimer structure is found (minimum C2 in Figure 71 and Figure 73).

These results suggest that the calcium ions located between EC1 and EC2 do in fact have an important role in the trans-swapping mechanism process. In fact, calcium ions seem to promote the arm opening by lowering the energy required for the closed-to-open transition, stabilizing the open form and also enhancing the sampling of monomer open conformations very similar to the X-ray swap-dimer structure.

Another interesting result comes from the analysis of the free energy surfaces projected on the CV1 and CV3 space (Figure 74 and Figure 75).

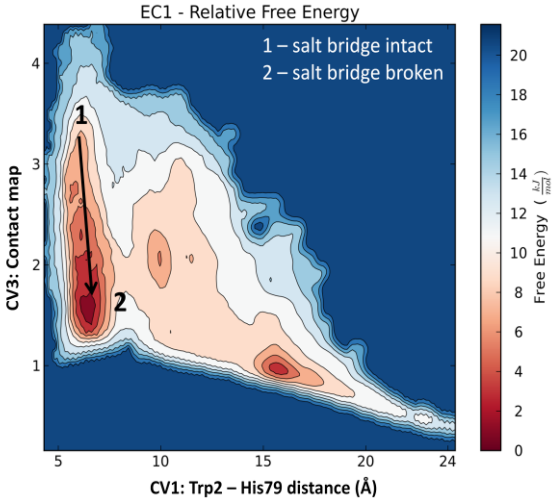


Figure 74. EC1 - Reconstructed Free Energy projected on a 2D surface using CV1 and CV3. Salt bridge refers to the electrostatic interaction between $NH_3^+$-Asp1 and the Glu89 side chain.

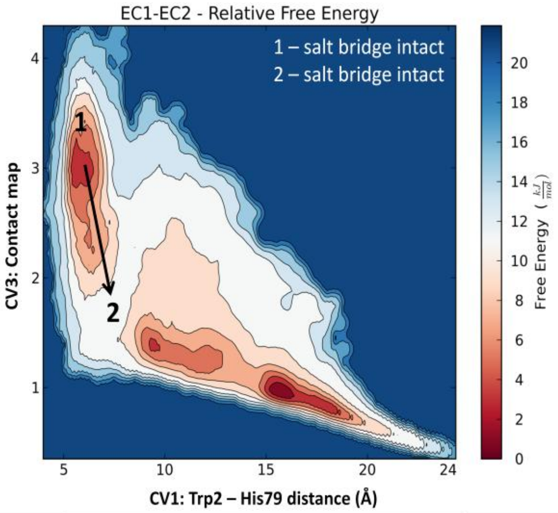


Figure 75. EC1-EC2 - Reconstructed Free Energy projected on a 2D surface using CV1 and CV3. Salt bridge refers to the electrostatic interaction between $NH_3^+$-Asp1 and the Glu89 side chain.

As mentioned in the introduction, it has been speculated that the adhesive arm, when in the closed conformation, acts as a strained "loaded spring", with two anchor points, one being the Trp2 docked into the hydrophobic pocket and other the $Ca^{2+}$ ions. The opening of the arm releases such a strain. By careful inspection of the 3D structures belonging to each basin in the reconstructed free energy surfaces, we could follow the opening path of both the EC1 and EC1-EC2 systems (from CV3=3, closed to CV3=1, open, Figure 75 and Figure 74), also highlighting the behavior of the $NH_3^+$-Asp1-Glu89 salt bridge in each basin.

In the case of EC1-EC2 we observed the expected mechanism: since the presence of calcium ions induces strain in the arm, we first saw the movement of the Trp2 side chain exiting the hydrophobic pocket (CV3=2). At this point, we noted that even though the indole moiety was exposed to the solvent, the arm was still bound to the rest of the protein by the $NH_3^+$-Asp1-Glu89 electrostatic interaction (Appendix D.3). Only in a second step the $NH_3^+$-Asp1-Glu89 salt bridge broke and the arm was then ready to fully open (corresponding to a drop in the CV3 value from 3 to 1). Furthermore, we observed many events in which, even though the salt bridge was still in place, the Trp2 side chain exited and reentered the hydrophobic pocket. As soon as the salt bridge vanished, these events became rare, suggesting that the breaking of this electrostatic interaction could be equally important to the opening mechanism. The opening of the arm for EC1-EC2 system could be summarized as follows:

1. Trp side chain leaves the binding pocket (CV3 from 3→2)
2. $NH_3^+$-Asp1-Glu89 electrostatic interaction is broken (CV3 from 2→1)

For EC1 system we observed that, by removing calcium ions, one of the two anchor points, the mechanism by which the conformational transition occurs changed. Removing the calcium ions at the end of EC1 released the strain and the $NH_3^+$-Asp1-Glu89 interaction became less relevant. First the salt bridge was broken (CV3=2, Figure 74) and the Trp2 side chain remained docked onto the pocket, having lost the driving force that facilitates the opening. (CV3=2). Thus, for EC1 system the path leading to the arm opening could be summarized as follows:

1. $NH_3^+$-Asp1-Glu89 electrostatic interaction is broken (CV3 from 3→2)
2. Trp side chain leaves the binding pocket (CV3 from 2→1)

This analysis confirms that the lack of calcium ions at the EC boundary deeply affects the mechanism of the arm opening in type I cadherins, not only by increasing the energy barrier for the closed-open transition and limiting the conformational space available to the open form, but also requiring a different mechanism opening path.

## 8.4 CONCLUSIONS

Cadherin dimerization mechanism is still vastly debated. Here, we proved that a sampling enhancing technique, like metadynamics, can be successfully used to gain insight into such an important process. The selected fit mechanism requires a first conformational transition leading to a fully open cadherin monomer, prior to the dimerization. We have confirmed the feasibility of this mechanism, and also highlighted the long range effect produced by the $Ca^{2+}$ ions.

However, more calculations are needed to test also the induced fit hypothesis. Even though the formation of the encounter complex lowers the activation energy for the trans dimerization, the path through which the encounter complex undergoes such a major conformational change is still completely unknown. We are confident that, with the use of metadynamics, we can also test this possibility, thus aiming at a full comprehension of the cadherin dimerization mechanism, that undoubtedly will help to design more selective and active inhibitors of the cadherin homophilic interaction.

## 8.5 METHODS

MD and MetaD simulations were performed using GROMACS 4.5.5[11] with the CHARMM22* force field[12] and PLUMED 1.3.[13] Initial closed structures were derived from the X-ray structure (pdb: 1ff5) of an E-cadherin X-dimer, in which the Trp2 side chain was docked in the hydrophobic pocket of the same cadherin.[7] Here, the crystallization of the E-cadherin in the X-dimer conformation was achieved by preceding the N-terminus with a methionine residue. This residue was then removed and the two monomeric closed systems were generated, one containing residues 1-99 (**EC1**) and no calcium ions, and the other containing residues 1-215 (**EC1-EC2**), three $Ca^{2+}$ ions at the EC boundary and a fourth calcium ion at the end of EC2. All the crystallographic waters were removed. **EC1** was solvated in a rhombic dodecahedron box adding 9253 TIP3P waters and **EC1-EC2** was solvated in a triclinic box adding 15136 TIP3P waters. Calcium ions were modeled following Bjelkmar and coworkers.[12] Neutralization was achieved by adding the necessary quantity of $Cl^-$ ions. The systems were energy minimized using a steepest descent algorithm. Then, a 1 ns equilibration at 300 K in an NVT ensemble was performed, using the Velocity-rescale thermostat[14] and positional restraint on the protein (k = 1000 kJ $mol^{-1}$ $nm^{-3}$), followed by 1 ns of NPT run with the same positional restraint on the protein, using a Parrinello-Rahman barostat.[15] Finally, 10 ns NPT simulation with no restraint concluded the equilibration step. Particle Mesh Ewald method[16] was used for treating long range electrostatics, using a cutoff of 10 Å. A time step of 2 fs was used for all simulations.

Preliminary WTE-PTMetaD runs were started by generating four exact replicas of each system, followed by a Parallel Tempering WTM run using the potential energy as the only CV. Temperatures for each replica are 300 K, 330 K, 362 K and 398 K, and were obtained from an exponential distribution $T_0 = T_i e^{k\,i}$, where k was set to be 0.094. Gaussians having height 4.0 kJ/mol were added every 500 MD steps, using a $\gamma$ of 120 (for **EC1**) and 220 (for **EC1-EC2**). The coordinates exchange among replicas was attempted every 250 MD steps.

Final WTE-PTMetaD runs were performed using the static biasing potential and using CV1, CV2 and CV3 as Collective variables. Gaussians of height 1.3 kJ/mol were added every 500 MD steps, using a $\gamma$ of 12 for both systems. The coordinates exchange among replicas was attempted every 250 MD steps. Simulations ran in NVT conditions with a time step of 2 fs until convergence (See Appendix D.4 for a graphical representation of the protocol being used).

## 8.6 BIBLIOGRAPHY

(1) Bennett, M. J.; Schlunegger, M. P.; Eisenberg, D. *Protein Sci.* **1995**, *4*, 2455–2468.

(2) Gronenborn, A. M. *Curr. Opin. Struct. Biol.* **2009**, *19*, 39–49.

(3) Vendome, J.; Posy, S.; Jin, X.; Bahna, F.; Ahlsen, G.; Shapiro, L.; Honig, B. *Nat. Struct. Mol. Biol.* **2011**, *18*, 693–700.

(4) Cailliez, F.; Lavery, R. *Biophys. J.* **2005**, *89*, 3895–903.

(5) Wu, Y.; Vendome, J.; Shapiro, L.; Ben-Shaul, A.; Honig, B. *Nature* **2011**, *475*, 510–513.

(6) D'Abramo, M.; Rabal, O.; Oyarzabal, J.; Gervasio, F. L. *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 642–646.

(7) Pertz, O.; Bozic, D.; Koch, a W.; Fauser, C.; Brancaccio, A.; Engel, J. *EMBO J.* **1999**, *18*, 1738–1747.

(8) http://www.plumed-code.org/documentation/manual_1-3-0.pdf.

(9) Deighan, M.; Pfaendtner, J. *Langmuir* **2013**, *29*, 7999–8009.

(10) Miloushev, V. Z.; Bahna, F.; Ciatto, C.; Ahlsen, G.; Honig, B.; Shapiro, L.; Palmer, A. G. *Structure* **2008**, *16*, 1195–1205.

(11) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(12) Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B.; Lindahl, E. *J. Chem. Theory Comput.* **2010**, *6*, 459–466.

(13) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. *Comput. Phys. Commun.* **2009**, *180*, 1961–1972.

(14) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101–014108.

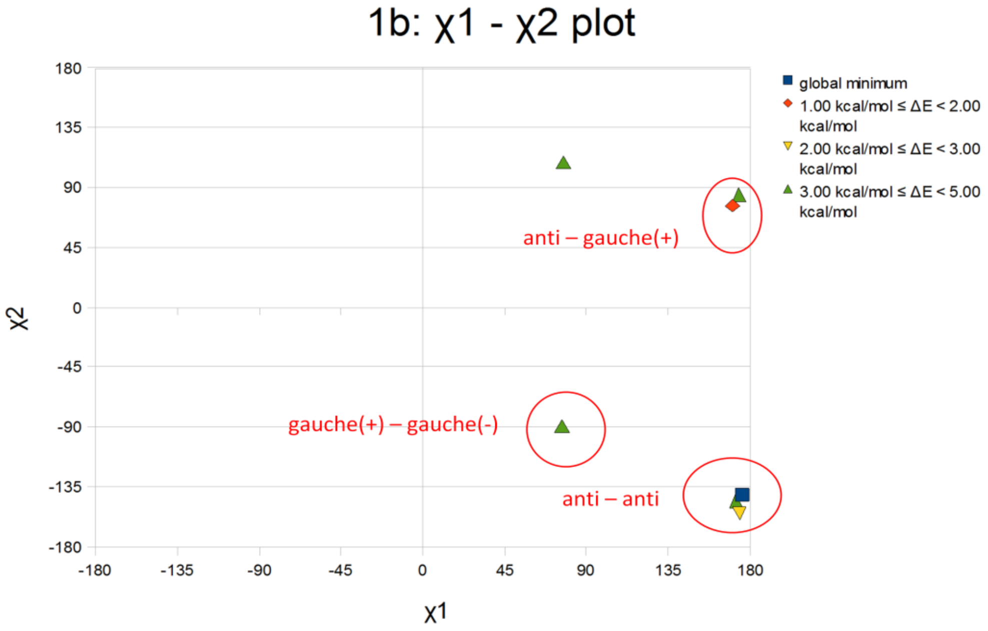
(15) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7185.

(16) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

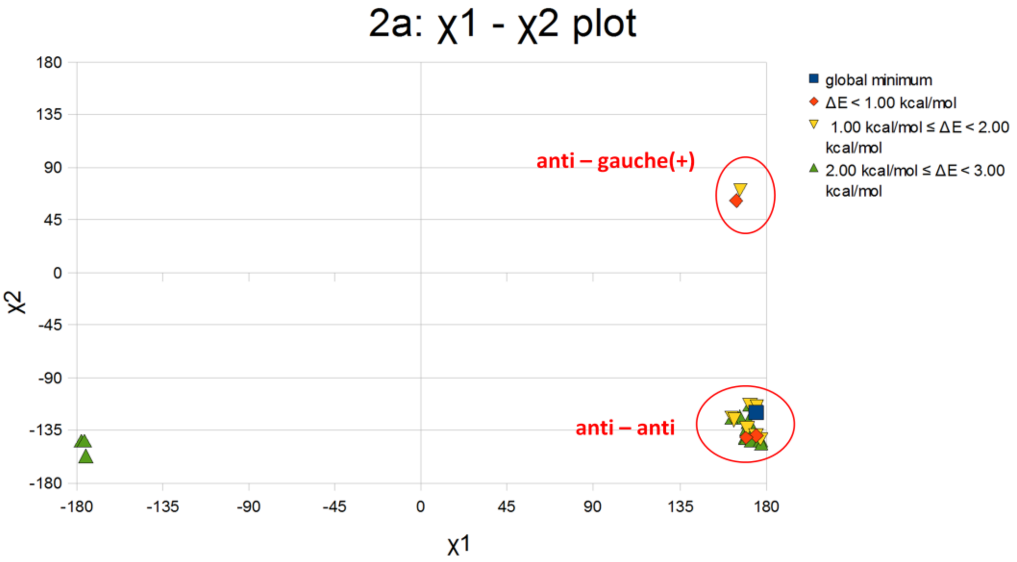
# Appendices

# A SUPPORTING INFORMATION FOR CHAPTER 4

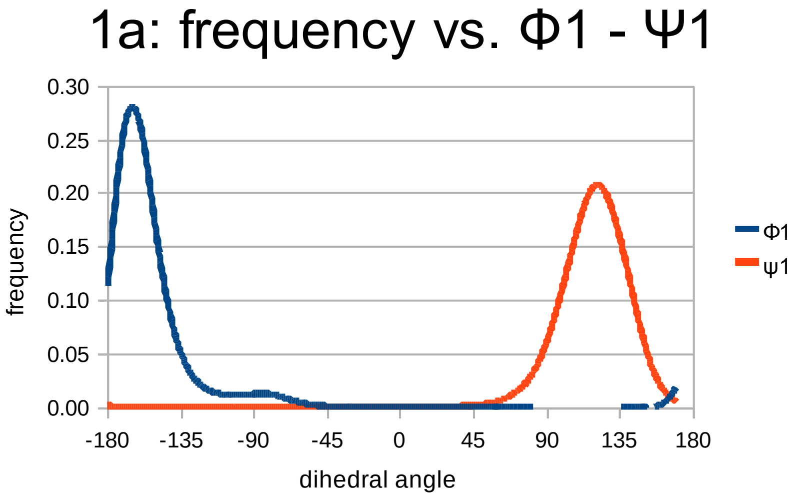
## A.1 $\chi^1$-$\chi^2$ plot for MC/EM of 1b.

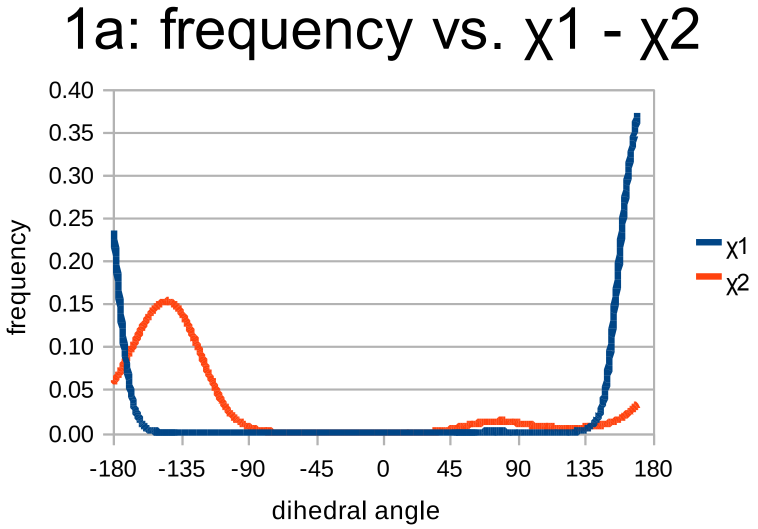


## A.2 $\chi^1$-$\chi^2$ plot for MC/EM of 2a.



### A.3 MC/SD dihedral angle distribution of 1a.

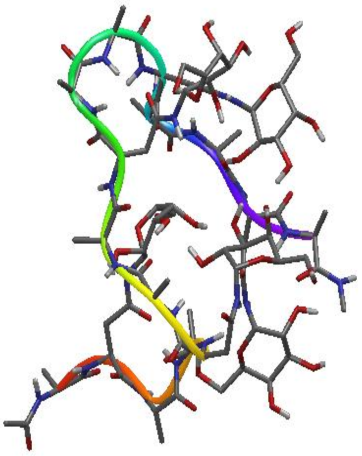




## A.4 Script to construct the 3D Ramachandran Plot for 3a and 4a.

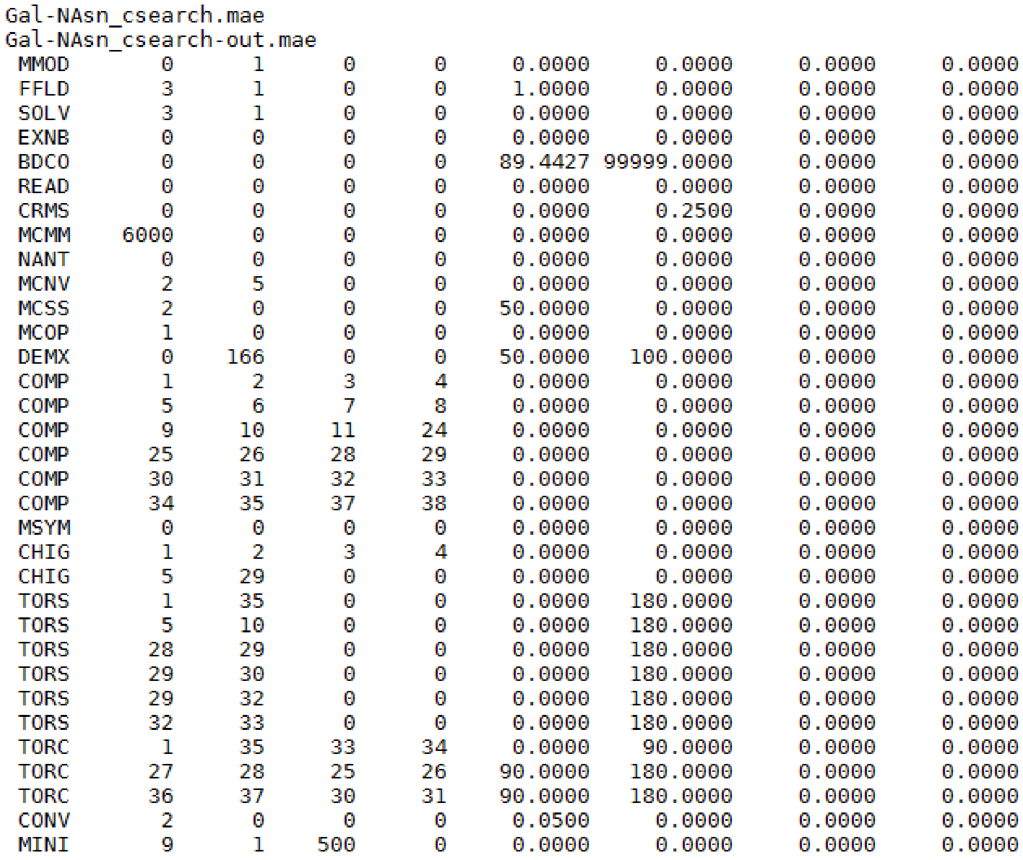
```python
from sys import argv
script_name, phi, psi = argv
file01=open(phi)
file02=open(psi)
Matrix={}
counter = 0
while True:
    counter += 1
    riga_phi=file01.readline()
    riga_psi=file02.readline()
    if riga_phi == '': break
    diedro01=int(float(riga_phi.split()[1]))
    diedro02=int(float(riga_psi.split()[1]))
    if (diedro01,diedro02) in Matrix:
        Matrix[(diedro01,diedro02)] += 1
    else:
        Matrix[(diedro01,diedro02)] = 1

file01.close()
file02.close()

output = open('output2.dat', 'w')

for i in range(-180,180,1):
    for j in range(-180,180,1):
        x=i+0.5
        y=j+0.5
        value = Matrix.get((i,j),0)
        output.write(str(x) + ' ' + str(y) + ' ' + str(value) + '\n' )
    output.write('\n')
output.close()
```

## A.5 Lowest energy structure in 50 ns simulation of 4a.



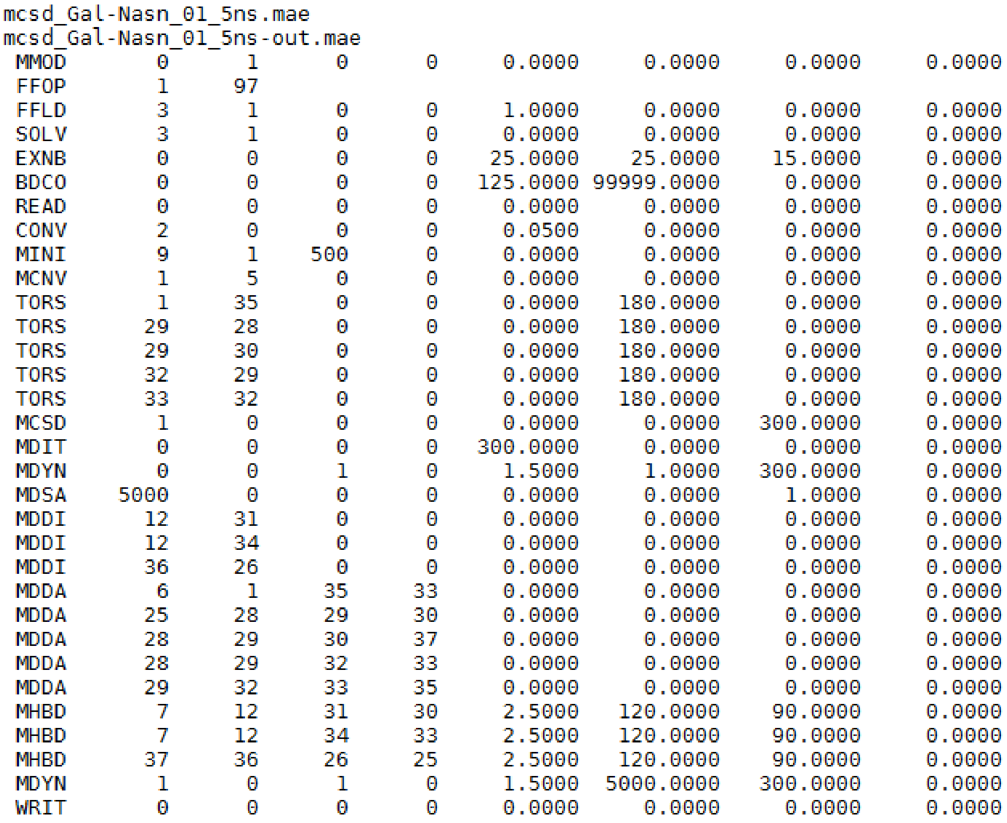
## A.6 MacroModel MC/EM and MC/SD command files.

MC/EM command file for **1a**



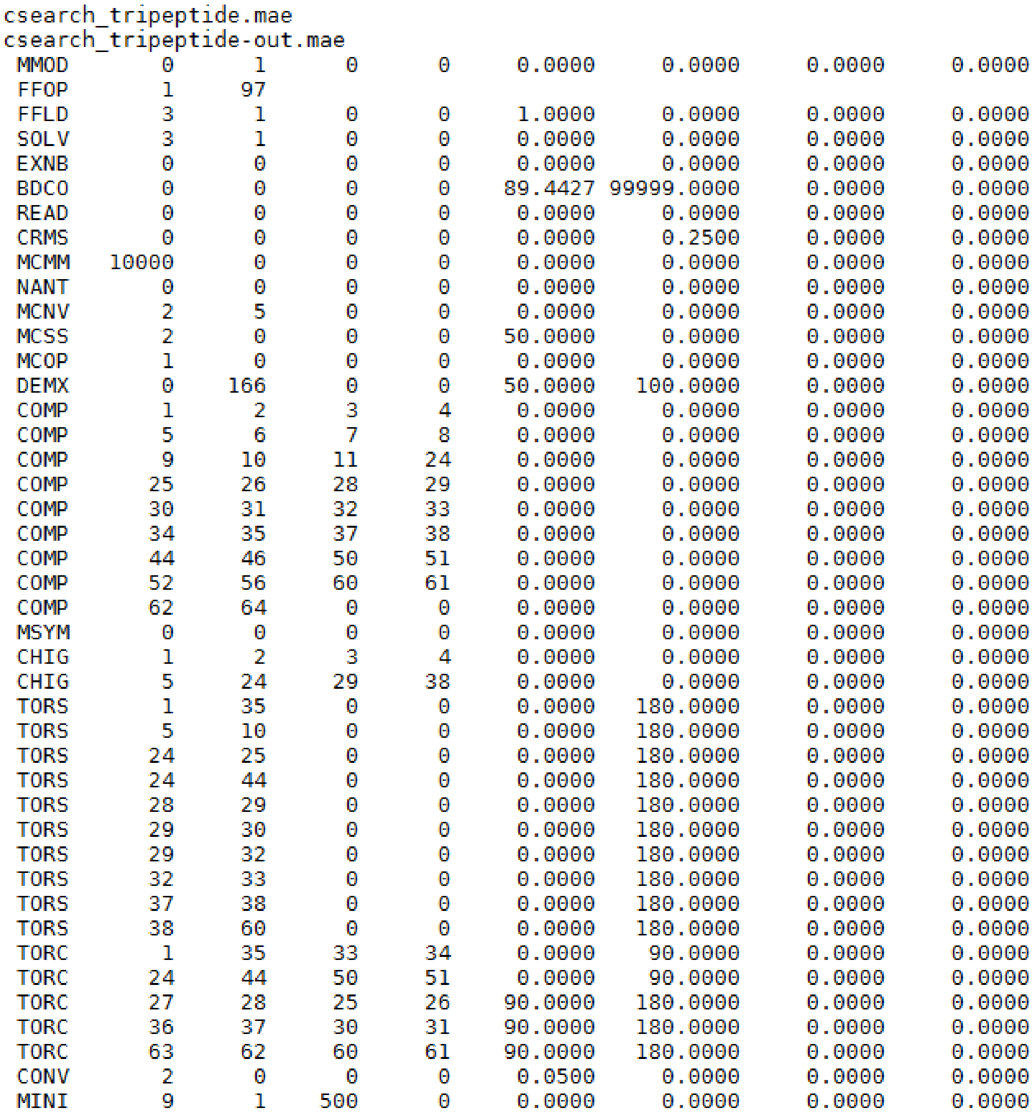
```
Gal-NAsn_csearch.mae
Gal-NAsn_csearch-out.mae
 MMOD       0      1      0      0     0.0000     0.0000     0.0000     0.0000
 FFLD       3      1      0      0     1.0000     0.0000     0.0000     0.0000
 SOLV       3      1      0      0     0.0000     0.0000     0.0000     0.0000
 EXNB       0      0      0      0     0.0000     0.0000     0.0000     0.0000
 BDCO       0      0      0      0    89.4427 99999.0000     0.0000     0.0000
 READ       0      0      0      0     0.0000     0.0000     0.0000     0.0000
 CRMS       0      0      0      0     0.0000     0.2500     0.0000     0.0000
 MCMM    6000      0      0      0     0.0000     0.0000     0.0000     0.0000
 NANT       0      0      0      0     0.0000     0.0000     0.0000     0.0000
 MCNV       2      5      0      0     0.0000     0.0000     0.0000     0.0000
 MCSS       2      0      0      0    50.0000     0.0000     0.0000     0.0000
 MCOP       1      0      0      0     0.0000     0.0000     0.0000     0.0000
 DEMX       0    166      0      0    50.0000   100.0000     0.0000     0.0000
 COMP       1      2      3      4     0.0000     0.0000     0.0000     0.0000
 COMP       5      6      7      8     0.0000     0.0000     0.0000     0.0000
 COMP       9     10     11     24     0.0000     0.0000     0.0000     0.0000
 COMP      25     26     28     29     0.0000     0.0000     0.0000     0.0000
 COMP      30     31     32     33     0.0000     0.0000     0.0000     0.0000
 COMP      34     35     37     38     0.0000     0.0000     0.0000     0.0000
 MSYM       0      0      0      0     0.0000     0.0000     0.0000     0.0000
 CHIG       1      2      3      4     0.0000     0.0000     0.0000     0.0000
 CHIG       5     29      0      0     0.0000     0.0000     0.0000     0.0000
 TORS       1     35      0      0     0.0000   180.0000     0.0000     0.0000
 TORS       5     10      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      28     29      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      29     30      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      29     32      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      32     33      0      0     0.0000   180.0000     0.0000     0.0000
 TORC       1     35     33     34     0.0000    90.0000     0.0000     0.0000
 TORC      27     28     25     26    90.0000   180.0000     0.0000     0.0000
 TORC      36     37     30     31    90.0000   180.0000     0.0000     0.0000
 CONV       2      0      0      0     0.0500     0.0000     0.0000     0.0000
 MINI       9      1    500      0     0.0000     0.0000     0.0000     0.0000
```

MC/SD command file for **1a**



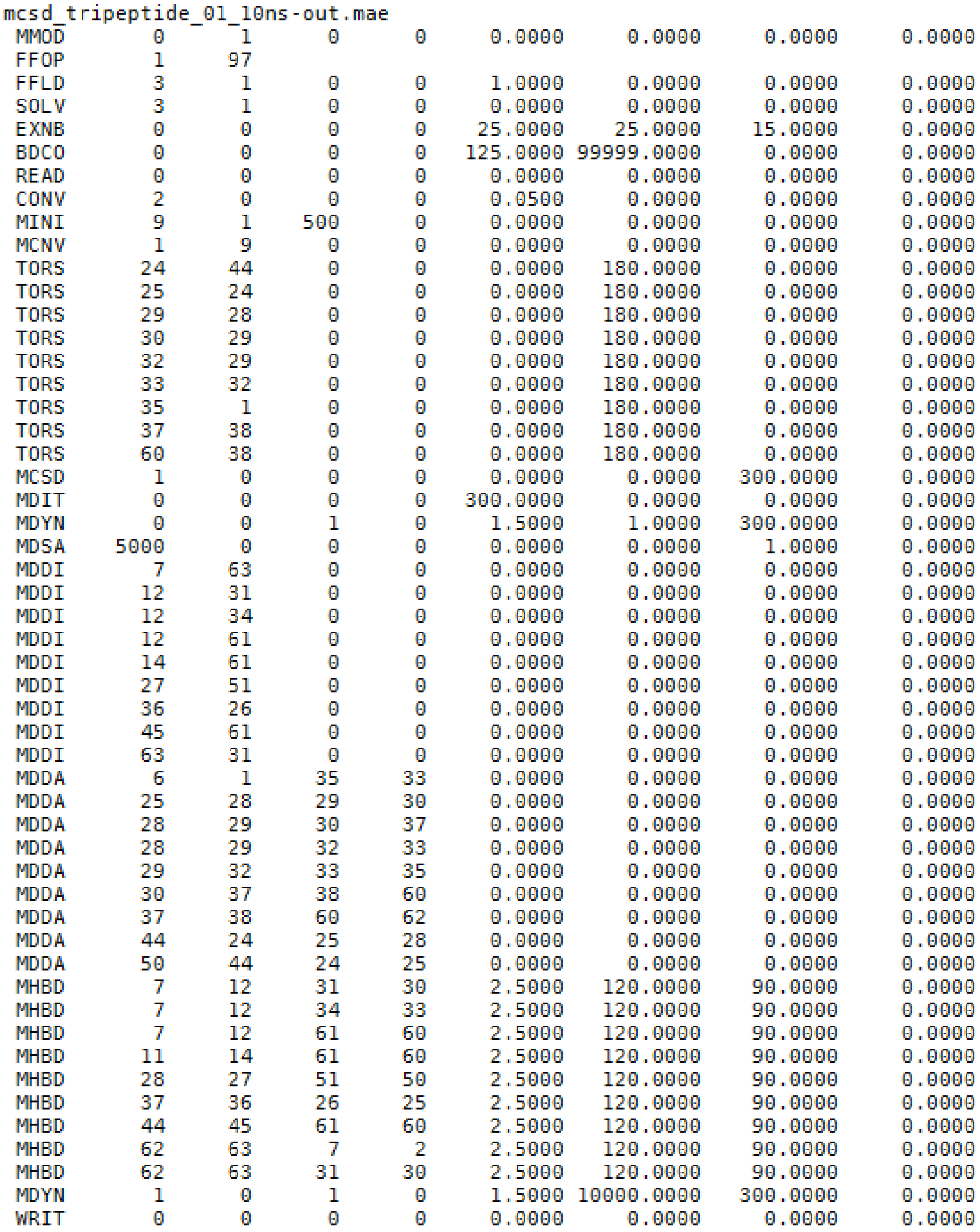
```
mcsd_Gal-Nasn_01_5ns.mae
mcsd_Gal-Nasn_01_5ns-out.mae
 MMOD       0      1      0      0     0.0000     0.0000     0.0000     0.0000
 FFOP       1     97
 FFLD       3      1      0      0     1.0000     0.0000     0.0000     0.0000
 SOLV       3      1      0      0     0.0000     0.0000     0.0000     0.0000
 EXNB       0      0      0      0    25.0000    25.0000    15.0000     0.0000
 BDCO       0      0      0      0   125.0000 99999.0000     0.0000     0.0000
 READ       0      0      0      0     0.0000     0.0000     0.0000     0.0000
 CONV       2      0      0      0     0.0500     0.0000     0.0000     0.0000
 MINI       9      1    500      0     0.0000     0.0000     0.0000     0.0000
 MCNV       1      5      0      0     0.0000     0.0000     0.0000     0.0000
 TORS       1     35      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      29     28      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      29     30      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      32     29      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      33     32      0      0     0.0000   180.0000     0.0000     0.0000
 MCSD       1      0      0      0     0.0000     0.0000   300.0000     0.0000
 MDIT       0      0      0      0   300.0000     0.0000     0.0000     0.0000
 MDYN       0      0      1      0     1.5000     1.0000   300.0000     0.0000
 MDSA    5000      0      0      0     0.0000     0.0000     1.0000     0.0000
 MDDI      12     31      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI      12     34      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI      36     26      0      0     0.0000     0.0000     0.0000     0.0000
 MDDA       6      1     35     33     0.0000     0.0000     0.0000     0.0000
 MDDA      25     28     29     30     0.0000     0.0000     0.0000     0.0000
 MDDA      28     29     30     37     0.0000     0.0000     0.0000     0.0000
 MDDA      28     29     32     33     0.0000     0.0000     0.0000     0.0000
 MDDA      29     32     33     35     0.0000     0.0000     0.0000     0.0000
 MHBD       7     12     31     30     2.5000   120.0000    90.0000     0.0000
 MHBD       7     12     34     33     2.5000   120.0000    90.0000     0.0000
 MHBD      37     36     26     25     2.5000   120.0000    90.0000     0.0000
 MDYN       1      0      1      0     1.5000  5000.0000   300.0000     0.0000
 WRIT       0      0      0      0     0.0000     0.0000     0.0000     0.0000
```

MC/EM command file for **2a**



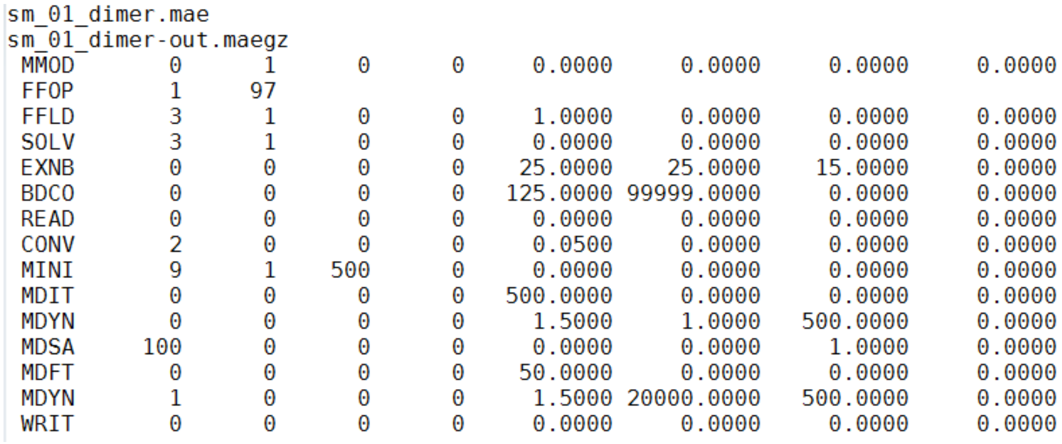
```
csearch_tripeptide.mae
csearch_tripeptide-out.mae
 MMOD      0      1      0      0     0.0000     0.0000     0.0000     0.0000
 FFOP      1     97
 FFLD      3      1      0      0     1.0000     0.0000     0.0000     0.0000
 SOLV      3      1      0      0     0.0000     0.0000     0.0000     0.0000
 EXNB      0      0      0      0     0.0000     0.0000     0.0000     0.0000
 BDCO      0      0      0      0    89.4427 99999.0000     0.0000     0.0000
 READ      0      0      0      0     0.0000     0.0000     0.0000     0.0000
 CRMS      0      0      0      0     0.0000     0.2500     0.0000     0.0000
 MCMM  10000      0      0      0     0.0000     0.0000     0.0000     0.0000
 NANT      0      0      0      0     0.0000     0.0000     0.0000     0.0000
 MCNV      2      5      0      0     0.0000     0.0000     0.0000     0.0000
 MCSS      2      0      0      0    50.0000     0.0000     0.0000     0.0000
 MCOP      1      0      0      0     0.0000     0.0000     0.0000     0.0000
 DEMX      0    166      0      0    50.0000   100.0000     0.0000     0.0000
 COMP      1      2      3      4     0.0000     0.0000     0.0000     0.0000
 COMP      5      6      7      8     0.0000     0.0000     0.0000     0.0000
 COMP      9     10     11     24     0.0000     0.0000     0.0000     0.0000
 COMP     25     26     28     29     0.0000     0.0000     0.0000     0.0000
 COMP     30     31     32     33     0.0000     0.0000     0.0000     0.0000
 COMP     34     35     37     38     0.0000     0.0000     0.0000     0.0000
 COMP     44     46     50     51     0.0000     0.0000     0.0000     0.0000
 COMP     52     56     60     61     0.0000     0.0000     0.0000     0.0000
 COMP     62     64      0      0     0.0000     0.0000     0.0000     0.0000
 MSYM      0      0      0      0     0.0000     0.0000     0.0000     0.0000
 CHIG      1      2      3      4     0.0000     0.0000     0.0000     0.0000
 CHIG      5     24     29     38     0.0000     0.0000     0.0000     0.0000
 TORS      1     35      0      0     0.0000   180.0000     0.0000     0.0000
 TORS      5     10      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     24     25      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     24     44      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     28     29      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     29     30      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     29     32      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     32     33      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     37     38      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     38     60      0      0     0.0000   180.0000     0.0000     0.0000
 TORC      1     35     33     34     0.0000    90.0000     0.0000     0.0000
 TORC     24     44     50     51     0.0000    90.0000     0.0000     0.0000
 TORC     27     28     25     26    90.0000   180.0000     0.0000     0.0000
 TORC     36     37     30     31    90.0000   180.0000     0.0000     0.0000
 TORC     63     62     60     61    90.0000   180.0000     0.0000     0.0000
 CONV      2      0      0      0     0.0500     0.0000     0.0000     0.0000
 MINI      9      1    500      0     0.0000     0.0000     0.0000     0.0000
```

MC/SD command file for **2a**

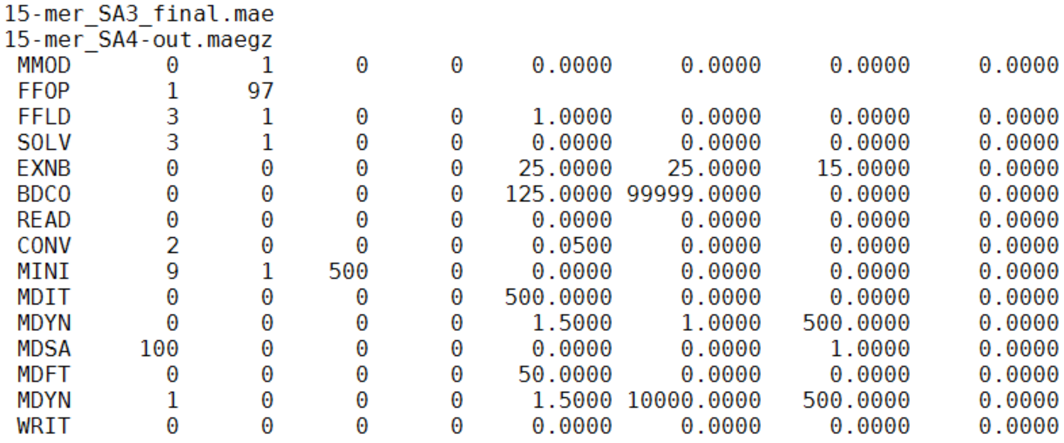


```
mcsd_tripeptide_01_10ns-out.mae
 MMOD      0      1      0      0     0.0000     0.0000     0.0000     0.0000
 FFOP      1     97
 FFLD      3      1      0      0     1.0000     0.0000     0.0000     0.0000
 SOLV      3      1      0      0     0.0000     0.0000     0.0000     0.0000
 EXNB      0      0      0      0    25.0000    25.0000    15.0000     0.0000
 BDCO      0      0      0      0   125.0000 99999.0000     0.0000     0.0000
 READ      0      0      0      0     0.0000     0.0000     0.0000     0.0000
 CONV      2      0      0      0     0.0500     0.0000     0.0000     0.0000
 MINI      9      1    500      0     0.0000     0.0000     0.0000     0.0000
 MCNV      1      9      0      0     0.0000     0.0000     0.0000     0.0000
 TORS     24     44      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     25     24      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     29     28      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     30     29      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     32     29      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     33     32      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     35      1      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     37     38      0      0     0.0000   180.0000     0.0000     0.0000
 TORS     60     38      0      0     0.0000   180.0000     0.0000     0.0000
 MCSD      1      0      0      0     0.0000     0.0000   300.0000     0.0000
 MDIT      0      0      0      0   300.0000     0.0000     0.0000     0.0000
 MDYN      0      0      1      0     1.5000     1.0000   300.0000     0.0000
 MDSA   5000      0      0      0     0.0000     0.0000     1.0000     0.0000
 MDDI      7     63      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI     12     31      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI     12     34      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI     12     61      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI     14     61      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI     27     51      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI     36     26      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI     45     61      0      0     0.0000     0.0000     0.0000     0.0000
 MDDI     63     31      0      0     0.0000     0.0000     0.0000     0.0000
 MDDA      6      1     35     33     0.0000     0.0000     0.0000     0.0000
 MDDA     25     28     29     30     0.0000     0.0000     0.0000     0.0000
 MDDA     28     29     30     37     0.0000     0.0000     0.0000     0.0000
 MDDA     28     29     32     33     0.0000     0.0000     0.0000     0.0000
 MDDA     29     32     33     35     0.0000     0.0000     0.0000     0.0000
 MDDA     30     37     38     60     0.0000     0.0000     0.0000     0.0000
 MDDA     37     38     60     62     0.0000     0.0000     0.0000     0.0000
 MDDA     44     24     25     28     0.0000     0.0000     0.0000     0.0000
 MDDA     50     44     24     25     0.0000     0.0000     0.0000     0.0000
 MHBD      7     12     31     30     2.5000   120.0000    90.0000     0.0000
 MHBD      7     12     34     33     2.5000   120.0000    90.0000     0.0000
 MHBD      7     12     61     60     2.5000   120.0000    90.0000     0.0000
 MHBD     11     14     61     60     2.5000   120.0000    90.0000     0.0000
 MHBD     28     27     51     50     2.5000   120.0000    90.0000     0.0000
 MHBD     37     36     26     25     2.5000   120.0000    90.0000     0.0000
 MHBD     44     45     61     60     2.5000   120.0000    90.0000     0.0000
 MHBD     62     63      7      2     2.5000   120.0000    90.0000     0.0000
 MHBD     62     63     31     30     2.5000   120.0000    90.0000     0.0000
 MDYN      1      0      1      0     1.5000 10000.0000   300.0000     0.0000
 WRIT      0      0      0      0     0.0000     0.0000     0.0000     0.0000
```

## A.7 MacroModel SA command files for 3a and 4a.



```
sm_01_dimer.mae
sm_01_dimer-out.maegz
 MMOD       0      1      0      0     0.0000     0.0000     0.0000     0.0000
 FFOP       1     97
 FFLD       3      1      0      0     1.0000     0.0000     0.0000     0.0000
 SOLV       3      1      0      0     0.0000     0.0000     0.0000     0.0000
 EXNB       0      0      0      0    25.0000    25.0000    15.0000     0.0000
 BDCO       0      0      0      0   125.0000 99999.0000     0.0000     0.0000
 READ       0      0      0      0     0.0000     0.0000     0.0000     0.0000
 CONV       2      0      0      0     0.0500     0.0000     0.0000     0.0000
 MINI       9      1    500      0     0.0000     0.0000     0.0000     0.0000
 MDIT       0      0      0      0   500.0000     0.0000     0.0000     0.0000
 MDYN       0      0      0      0     1.5000     1.0000   500.0000     0.0000
 MDSA     100      0      0      0     0.0000     0.0000     1.0000     0.0000
 MDFT       0      0      0      0    50.0000     0.0000     0.0000     0.0000
 MDYN       1      0      0      0     1.5000 20000.0000   500.0000     0.0000
 WRIT       0      0      0      0     0.0000     0.0000     0.0000     0.0000
```

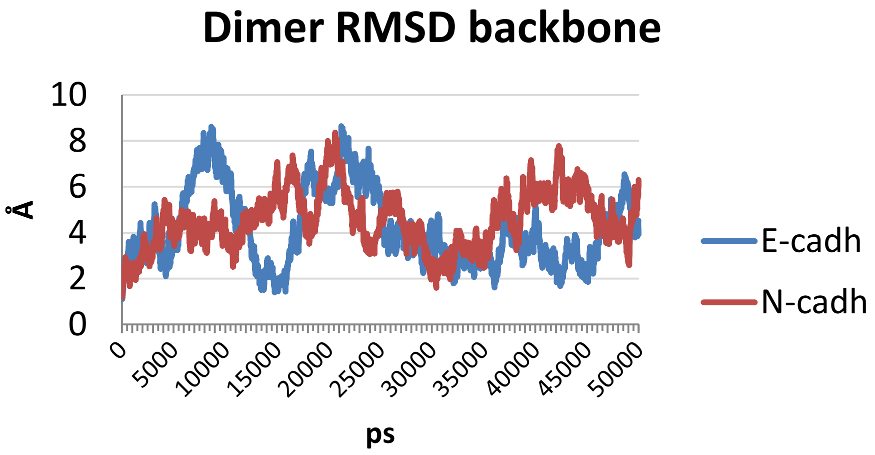


```
15-mer_SA3_final.mae
15-mer_SA4-out.maegz
 MMOD       0      1      0      0     0.0000     0.0000     0.0000     0.0000
 FFOP       1     97
 FFLD       3      1      0      0     1.0000     0.0000     0.0000     0.0000
 SOLV       3      1      0      0     0.0000     0.0000     0.0000     0.0000
 EXNB       0      0      0      0    25.0000    25.0000    15.0000     0.0000
 BDCO       0      0      0      0   125.0000 99999.0000     0.0000     0.0000
 READ       0      0      0      0     0.0000     0.0000     0.0000     0.0000
 CONV       2      0      0      0     0.0500     0.0000     0.0000     0.0000
 MINI       9      1    500      0     0.0000     0.0000     0.0000     0.0000
 MDIT       0      0      0      0   500.0000     0.0000     0.0000     0.0000
 MDYN       0      0      0      0     1.5000     1.0000   500.0000     0.0000
 MDSA     100      0      0      0     0.0000     0.0000     1.0000     0.0000
 MDFT       0      0      0      0    50.0000     0.0000     0.0000     0.0000
 MDYN       1      0      0      0     1.5000 10000.0000   500.0000     0.0000
 WRIT       0      0      0      0     0.0000     0.0000     0.0000     0.0000
```

# B SUPPORTING INFORMATION FOR CHAPTER 6

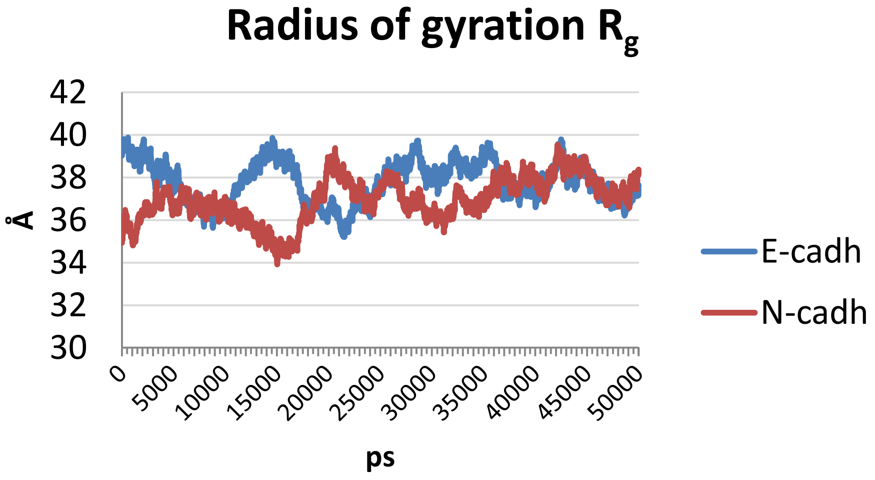
## B.1 RMSD values during the MD simulation

Time evolution of the dimer backbone (atoms Cα, C, N) RMSD for E- and N-cadherin during 50ns of MD.



## B.2 Radius of gyration during the MD simulation

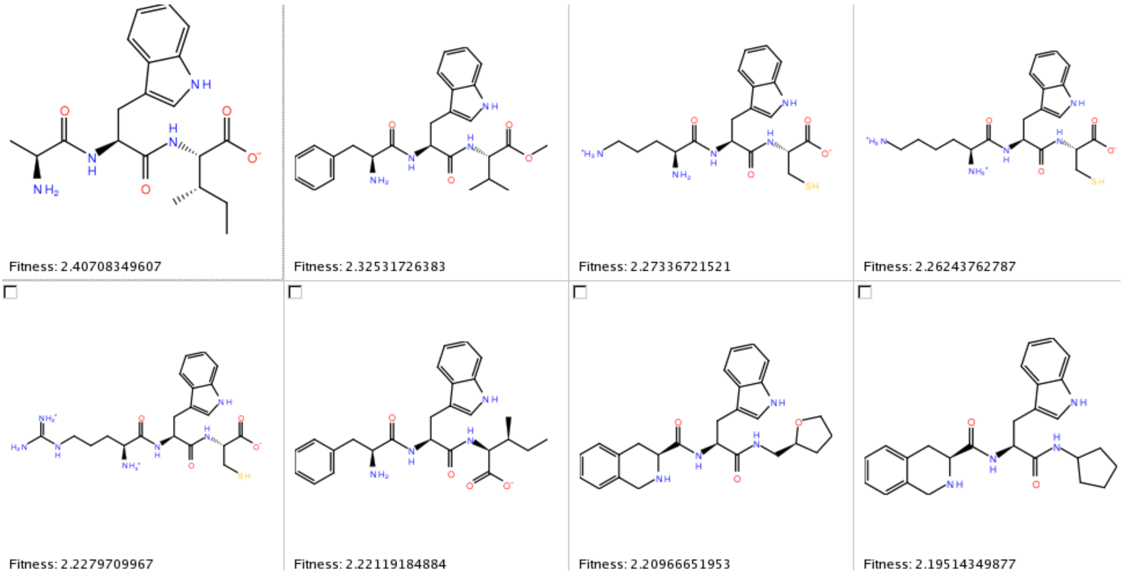
Time evolution of radius of gyration during the 50ns MD run of E-and N-cadherin dimers.



# C SUPPORTING INFORMATION FOR CHAPTER 7

## C.1 Top compounds extracted from PubChem

First 8 compounds out of 200 that satisfy the 5-site 3D pharmacophoric query applied on 37738 hits coming from PubChem. A fitness value of 3 corresponds to a perfect match.



Fitness: 2.40708349607

Fitness: 2.32531726383

Fitness: 2.27336721521

Fitness: 2.26243762787

Fitness: 2.2279709967

Fitness: 2.22119184884

Fitness: 2.20966651953

Fitness: 2.19514349877

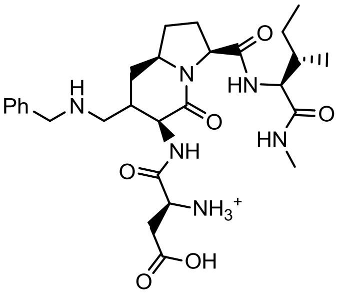
## C.2 Script used to filter PepMMsMimic results

Python script used to filter the PepMMsMimic dataset based on the SMILES sequence of each compound.

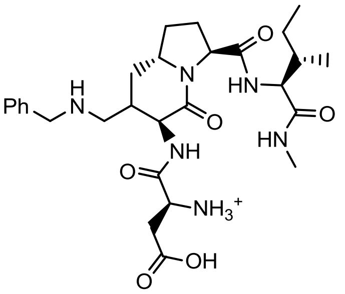
```
#!/usr/bin/env python
import sys
vocab = {}
elenco = set()
finalset = set()
x = len(sys.argv[1:])
#print x
for nome in sys.argv[1:]:
  for line in open(nome):
    temp1 = line.split()[0]
    temp2 = temp1.replace('"', '')
    #print temp2
    elenco.add(temp2)
  vocab[nome] = elenco
  vocab[nome].remove('Title')
  elenco = set()
for i in range(1,x):
  #print i
  if not finalset:
    a = vocab[sys.argv[i]]
    b = vocab[sys.argv[i+1]]
    finalset = set(a) & set(b)
  else:
    b = vocab[sys.argv[i+1]]
    finalset = finalset & set(b)
num = str(len(finalset))
results=open('results.log', 'w')
results.write('Files in input: ')
for i in range(1,x+1):
  results.write(sys.argv[i]+'\t')
results.write('\n')
results.write('\n')
results.write('Numero di strutture comuni: ' + num + '\n' )
results.write('Entries comuni ai file dati in input:'+ '\n')
for i in finalset:
  results.write(i)
  results.write('\n')
results.close()
print ('OK. I risultati sono in results.log')
```

## C.3 Virtual library of rationally designed compounds

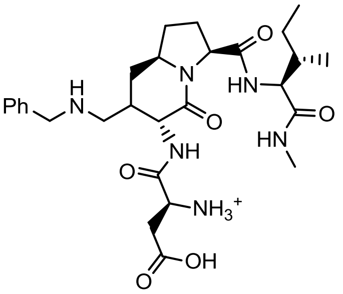
Virtual library designed and in silico screened of general formula $NH_3^+$-Asp-scaffold-Ile-$NHCH_3$. Molecules are ordered based on their scaffold (Figure 58, Chapter 7), from I to VI. Only compounds synthetically accessible were chosen to be in silico screened.



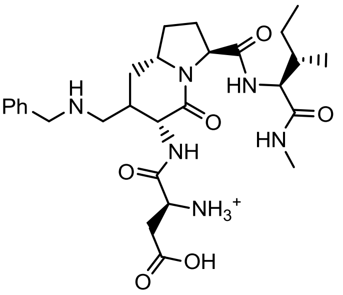
6-5b_001



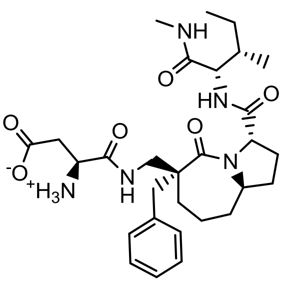
6-5b_002



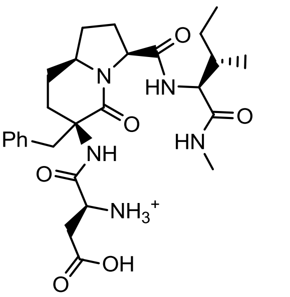
6-5b_003



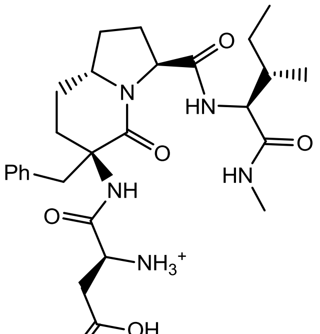
6-5b_004



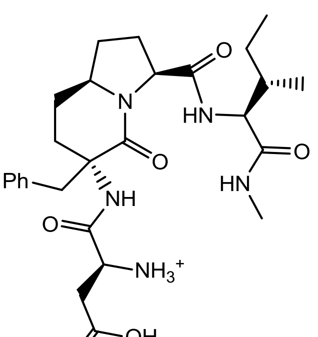
7-5trans_benzylate



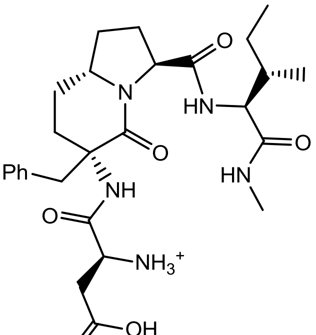
6-5_001



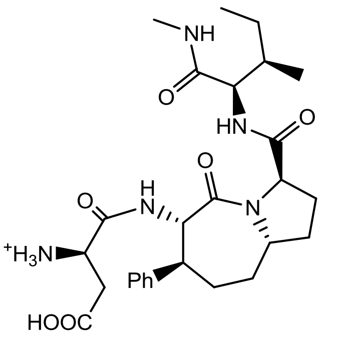
6-5_002



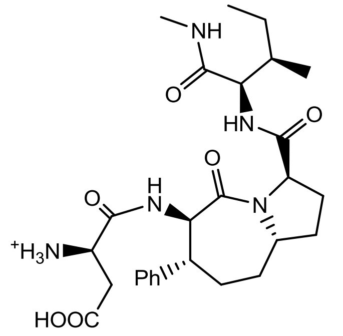
6-5_003



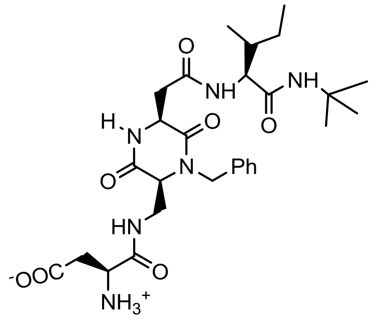
6-5_004



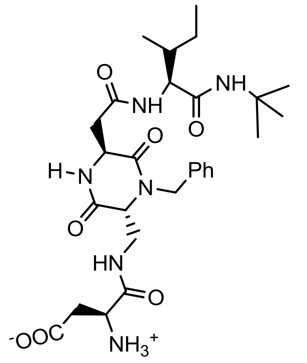
7-5trans_R-R_C-beta



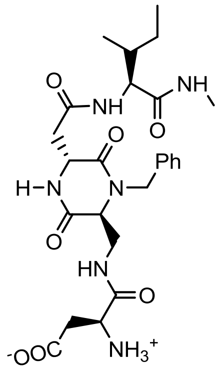
7-5trans_S-S_C-beta



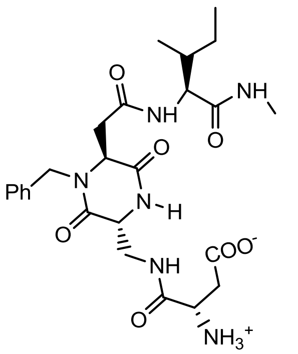
ASP-DKP1-ILE-NHMe
(3)



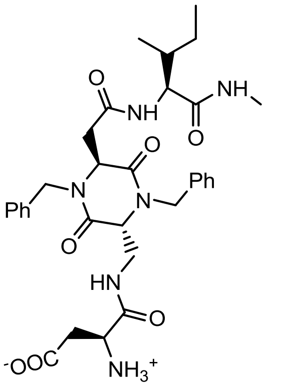
ASP-DKP2-ILE-NHMe
(2)



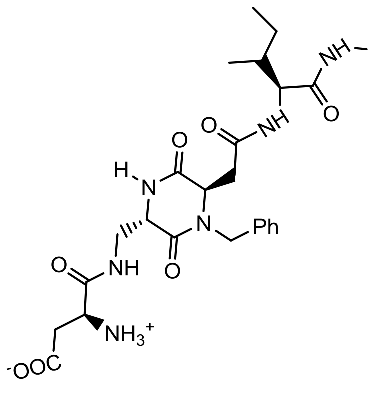
ASP-DKP3-ILE-NHMe



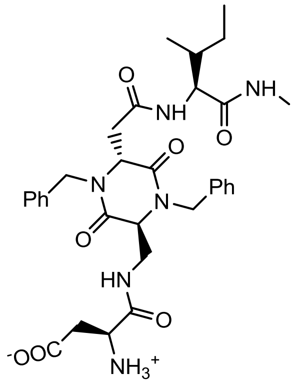
ASP-DKP4-ILE-NHMe



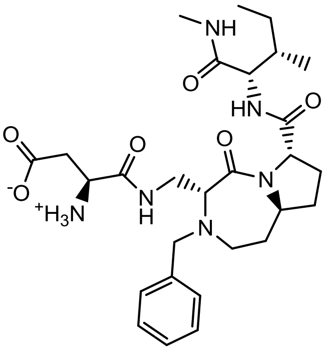
ASP-DKP5-ILE-NHMe



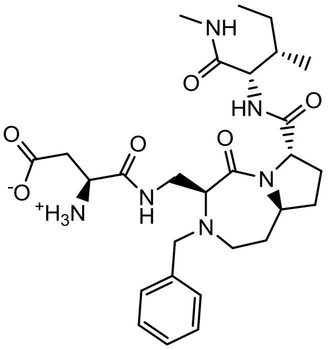
ASP-DKP6-ILE-NHMe



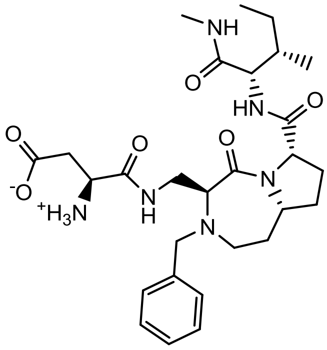
ASP-DKP7-ILE-NHMe



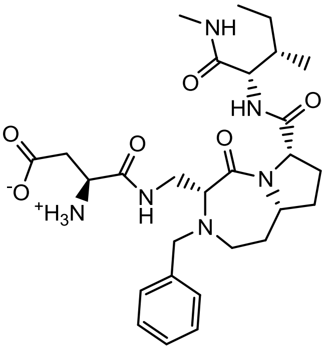
7-5_S-R_trans



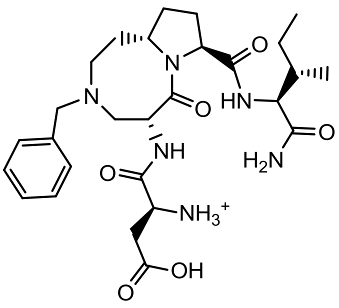
7-5_S-S_trans



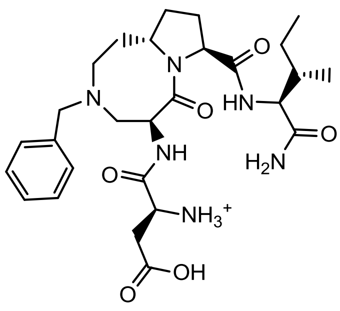
7-5cis_S-S



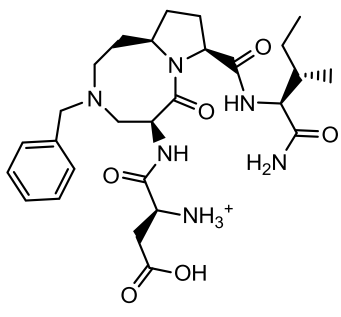
7-5cis_S-R



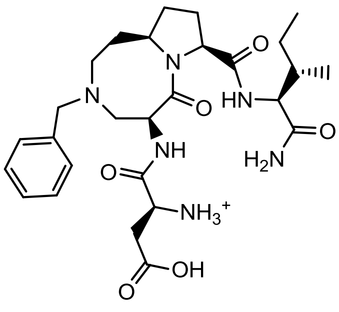
8-5_S-R_trans
(**1**)



8-5_S-S_trans



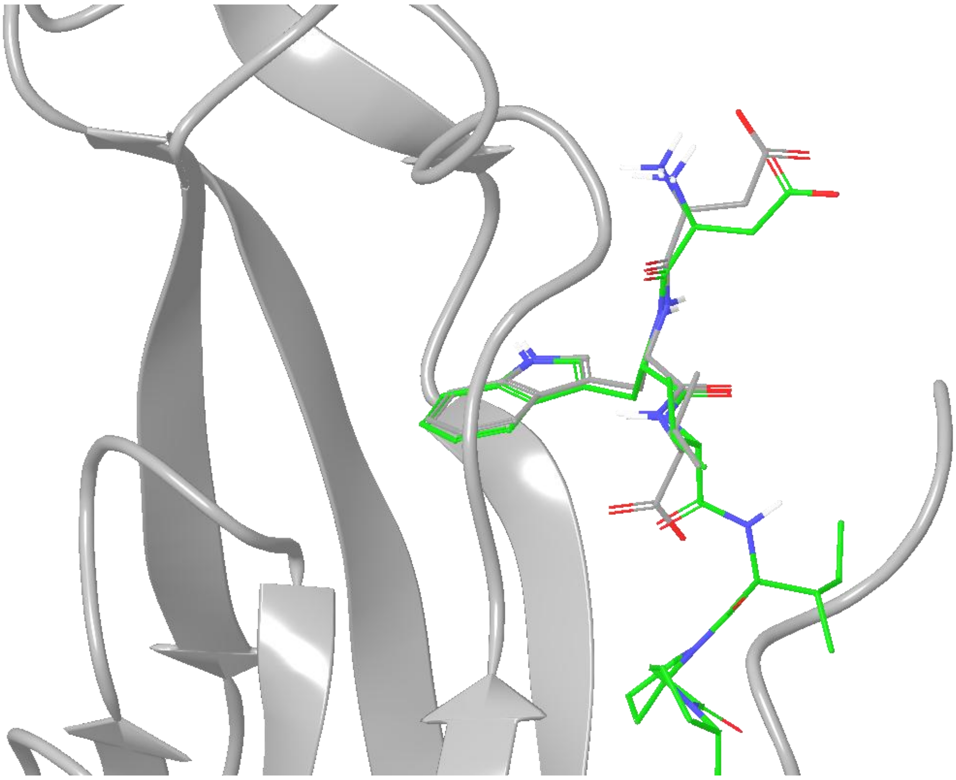
8-5_R-S_trans



8-5_R-R_trans

## C.4 DWV docked into N-cadherin EC1

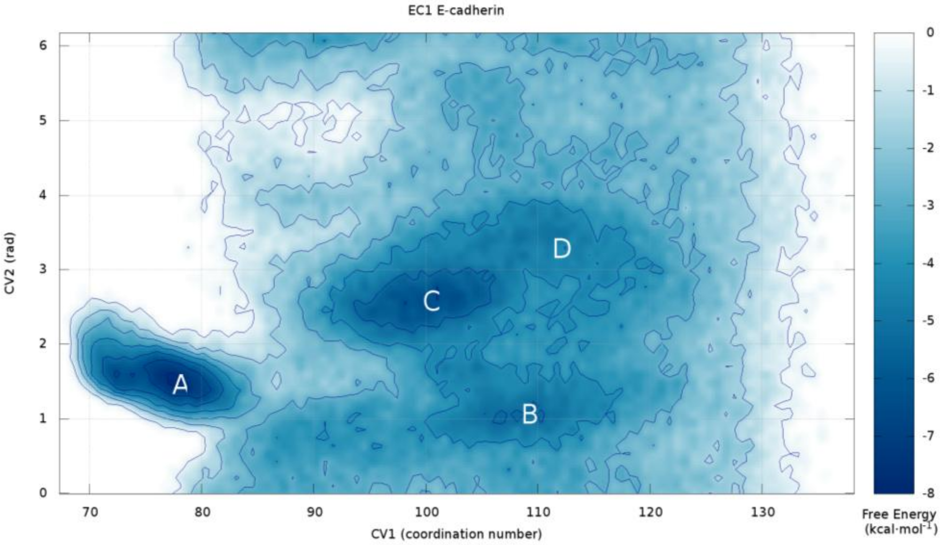
Superposition between the best pose of the tripeptide DWV (grey) docked into the N-cadherin EC1 (showed in ribbons) and the corresponding crystallographic N-terminus adhesive arm (green). RMSD computed on the heavy atoms is 0.7 Å.



# D SUPPORTING INFORMATION FOR CHAPTER 8

## D.1 FES reconstructed using CV1 and coordination number

Reconstructed free energy surface using CV1 (distance Trp2 - His79) and the coordination number (CV2 in the Figure). Minima from A to D represent conformations having the Trp2 indole ring hindered from the solvent (A) or variously exposed to the water (B-D). The use of the coordination number as CV was attempted by performing ca. 270 ns of Well Tempered Metadynamics (using a bias factor γ of 12), with no replica exchange. The simulation did not reach convergence.



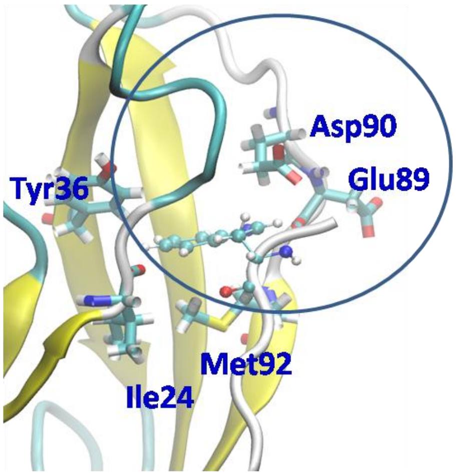
## D.2 Script used to check the MetaD convergence

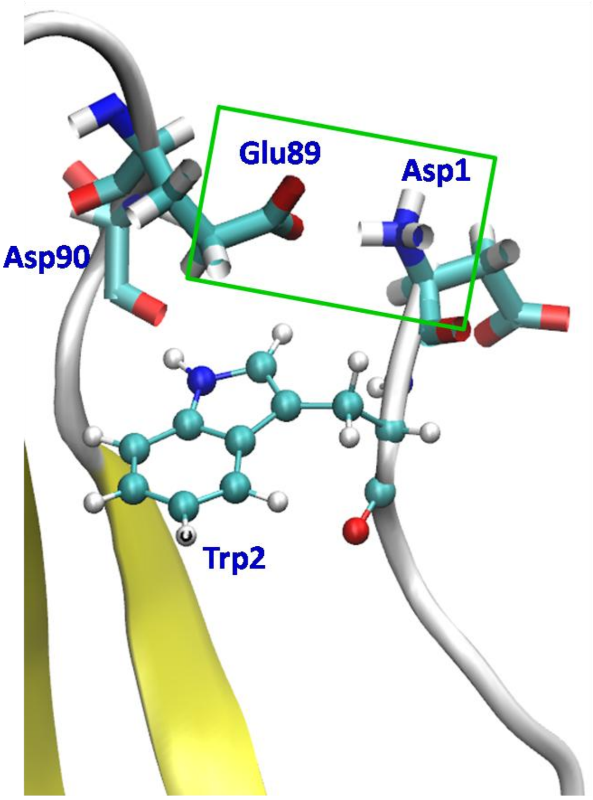
```python
#!/usr/bin/env python
"""
Compute minima differences in monodimensional fes
usage: conv_check.py x_iniziale x_finale
output: delta.dat which contains
fes number, minimum, difference to preceding minimum
"""
import os
import numpy as np
from sys import argv

x_i = float(argv[1])
x_f = float(argv[2])
minimi = []
deltas = []

lista_file = os.listdir(os.getcwd())

dat_files = [ i for i in lista_file if i.startswith('fes.dat.') ]
#put zero to first 9 files, and sort
for index, dat in enumerate(dat_files):
    if len(dat[8:]) == 1:
        new = dat[:8] + '0' + dat[-1]
        dat_files[index] = new

dat_files.sort()

#Remove zero
for i in range(9):
    dat_files[i]= dat_files[i][:8] + dat_files[i][-1]

for i, dat in enumerate(dat_files):
    x,y = np.loadtxt(dat, unpack=True)
    cond = np.logical_and(x > x_i, x < x_f)
    minimo = min(y[cond])
    minimi.append(minimo)
    try:
        delta = abs(minimi[i] - minimi[i-1])
    except:
        delta = 0

    deltas.append(delta)

line = '{0:1d}\t{1:.3f}\t{2:.3f}\n'
dat = range(1,len(dat_files)+1)
with open('delta.dat', 'w') as results:
    for a,b,c in zip(dat,minimi, deltas):
        results.write(line.format(a,b,c))
```

## D.3 E-cadherin closed form

Side chain of E-cadherin Trp2 docked inside its own hydrophobic pocket (structure generated from the modified X-dimer structure, pdb code 1ff5).

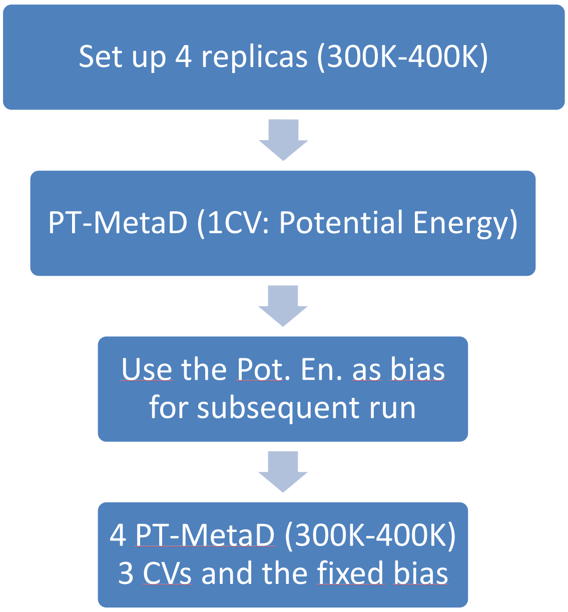
To the left, an overview of the residues interacting with the adhesive arm can be seen. To the right, the salt bridge formed between the side chain of Glu89 and the N-terminus $NH_3^+$ of Asp1 is highlighted.





### D.4 WTE-PTMetaD protocol

Protocol used to apply the Parallel Tempering Metadynamics in the Well Tempered Ensemble algorithm to the E-cadherin systems.



Set up 4 replicas (300K-400K)

PT-MetaD (1CV: Potential Energy)

Use the Pot. En. as bias for subsequent run

4 PT-MetaD (300K-400K) 3 CVs and the fixed bias