In Silico Studies on DARC.
Alexandre de Brevern, Ludovic Autin, yves Colin, Olivier Bertrand, Catherine Etchebest

To cite this version:

HAL Id: inserm-00366309
https://www.hal.inserm.fr/inserm-00366309
Submitted on 1 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
To be published in Infectious Diseases – Drug Targets (2009)

in silico studies on DARC

Alexandre G. de Brevern*, Ludovic Autin#, Yves Colin, Olivier Bertrand & Catherine Etchebest

INSERM UMR-S 665, Université Paris Diderot – Paris 7, Institut National de la Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris 15, France.

Short title: in silico DARC

* Corresponding author:
mailing address: Dr. de Brevern A.G., INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Institut National de la Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris 15, France
E-mail: alexandre.debrevern@univ-paris-diderot.fr
Tel: (33) 1 44 49 30 38
Fax: (33) 1 43 06 50 19

# present address : Molecular Graphics Laboratory, Department of Molecular Biology, MB-5, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037-1000, USA.
Abstract

The Duffy Antigen/Receptor for Chemokine (DARC) is a seven segment transmembrane protein. It was firstly discovered as a blood group antigen and was the first specific gene locus assigned to a specific autosome in man. It became more famous as an erythrocyte receptor for malaria parasites (*Plasmodium vivax* and *Plasmodium knowlesi*), and finally for chemokines. DARC is an unorthodox chemokine receptor as (i) it binds chemokines of both CC and CXC classes and (ii) it lacks the Asp-Arg-Tyr consensus motif in its second cytoplasmic loop hence cannot couple to G proteins and activate their signaling pathways. DARC had also been associated to cancer progression, numerous inflammatory diseases, and possibly to AIDS.

In this review, we will summarize important biological data on DARC. Then we shall focus on recent development of the elaboration and analyzes of structural models of DARC. We underline the difficulty to propose pertinent structural models of transmembrane protein using comparative modeling process, and other dedicated approaches as the Protein Blocks. The chosen structural models encompass most of the biochemical data known to date. Finally, we present recent development of protein–protein docking between DARC structural models and CXCL-8 structures. We propose a hierarchal search based on separated rigid and flexible docking.

**Key-words:** Duffy Antigen / Receptor for Chemokine; CXCL-8, Duffy Binding Protein, chemokines, malaria, *Plasmodium vivax*. 
Biology and Pathology of the DARC

The Duffy antigen receptor for chemokines (DARC) is a protein which has a long history. It was discovered at the beginning of the 50’s as a blood group antigen, after the identification of antibodies against the protein in the serum of a polytransfused haemophilic named Duffy. DARC carries the antithetic antigens Fy\textsuperscript{a} and Fy\textsuperscript{b} in humans [1, 2]. DARC, as a blood group, was extensively studied and its mode of inheritance defined [3-12]. Interestingly, it was the first specific gene locus assigned to a specific autosome in man [13].

A milestone in DARC studies was the discovery that it is an erythrocyte receptor for malaria parasites. Indeed, \textit{in vitro} and \textit{in vivo} studies performed on American volunteering detainees have shown that erythrocytes of Fy(a\textsuperscript{b}) individuals that do not express Duffy cannot be invaded by \textit{Plasmodium knowlesi} [14] and \textit{Plasmodium vivax} [15]. The Duffy negative phenotype is due to homozygosity for a promoter polymorphism (-46C) that disrupts the binding site for a transcription factor (GATA-1) required for the DARC chemokine receptor to be expressed on the cell surface of erythrocytes [16]. This mutation is ubiquitous in West Africa proving that it was due to a strong selection pressure. An important point to be stressed is that this mutation abolishes expression of DARC in the erythrocytic lineage but does not hinder expression in other cell types in which another promoter is operative.

It must be noticed that \textit{Plasmodium vivax} malaria [17] is the most geographically widespread and the second prevalent cause of malaria. Among 2.6 billion people at risk of malaria infection, annual estimates of \textit{P. vivax} cases range from 130 to 435 million [18]. 90% of these infections occur outside Africa [19]. \textit{P. vivax} transmission agents to humans are \textit{Anopheles} of various species [20]. Chloroquine was one of the first and still the most commonly used drug for treatment for \textit{P. vivax} malaria. However, chloroquine resistance \textit{P. vivax} becomes a threatening problem [21-29].

Searching for a vaccine eliciting antibodies able to interfere with interaction of DARC...
with its malarial partner the Duffy Binding Protein (or DBP) is an active research field [30-33]. However some recent data were issued in the literature suggesting that *P. vivax* might infest Duffy negative people in East Africa and Brazil [34-37].

DARC, while it was well known as a receptor for *P. vivax* was more recently identified as a chemokine receptor because Duffy-positive but not Duffy-null erythrocytes can bind CXCL8 (IL-8) [38, 39] and is involved in physiological chemiotaxis [40]. Contrary to other chemokine receptors, it binds members of both the CC and CXC chemokine families. It binds angiogenic ELR\(^+\) (*i.e.* with the Glutamine–Leucine–Arginine motif at the N-terminus) CXC chemokines, as well as some CC chemokines [41-43], but it does not bind ELR\(^-\) CXC chemokines and the C chemokine [44]. DARC ligands with high affinity include CXCL1 (GRO\(\alpha\)), CXCL2 (GRO\(\beta\)), CXCL3 (GRO\(\gamma\)), CXCL4 (PF4), CXCL5 (ENA-78), CXCL7 (NAP-2), CXCL8 (IL-8), CCL2 (MCP-1, MCAF, JE), CCL5 (RANTES), CCL7 (MCP-3), CCL11 (Eotaxin), CCL13 (MCP-4), CCL14 (HCC-1) and CCL17 (TARC) [45].

The DARC receptor plays a complex role in regulating chemokine levels, at least in part due to its ability to bind chemokines and adsorb them onto the red cell surface, thereby acting as a chemokine reservoir; it is so considered as typical decoy receptor or a sink [38]. Chemokines binding to DARC can largely be categorized as inflammatory. Chemokine binding to DARC does not lead to intracellular signaling. Because DARC lacks the Asp-Arg-Tyr consensus motif in its second cytoplasmic loop, it cannot couple to G proteins and activate the subsequent signaling pathways [46].

DARC was associated to numerous physiological and pathological responses [47] as the other atypical chemokine receptor D6 [48]. DARC is thought to play a part in homeostasis either through clearance of inflammatory chemokines from the circulation or maintenance of plasma levels by subsequent release from the red cell surface [49]. But the mechanisms by which DARC is implicated in inflammation are far to be completely understood. It has been
reported that DARC might heterodimerize with CXCR-5 and so interfere with function of this important effector [50].

DARC expressed in endothelial cells of post capillary venules might assist leukocyte migration to sites of inflammation [46], Noticeably it has been suggested that DARC might ensure transcytosis of chemokines from the intravascular to the extravascular spaces hence favor migration of leucocytes involved in inflammation [51]. This mechanism clearly is operative even in individuals lacking Duffy antigen expression on red cells. Interactions of chemokines with DARC support their activity on apposing leukocytes in vitro and in vivo [49]. The up-regulation of DARC expression by endothelial cells has been demonstrated in many human inflammatory and infectious diseases [52-59]. For instance, DARC is implicated in numerous pathologies implicating “classical” chemokine receptors like in giant cell arteritis [52]. Besides its role in inflammation, definitely DARC seems to play a role in cancer. Epidemiological data suggest that lack of DARC on erythrocytes favors early onset and aggressiveness of prostate cancer observed in African Americans [60]. These data are backed up by experimental models using genetically modified mice [61]. On the other hand expression of DARC on malignant lung or breast cancer cells might be associated with less growth and metastatic potential this being possibly related to sequestration of angiogenic chemokines and inhibition of tumor vascularisation [62, 63]. Susceptibility to asthma and atopy among certain populations of African descent was correlated also to absence of DARC expression on red cells [64].

A recent study suggests that DARC influences HIV/AIDS susceptibility by mediating trans-infection of HIV-1 and by affecting both chemokine-HIV interactions and chemokine-driven inflammation [65]. The almost universal presence of the HIV-susceptible DARC genotype in Africans might be an important contribution to understanding the massive extent of the HIV-1 epidemic in Africa [66, 67]. Worth of notice This result was considered as one
of the eight most important discovery of the year in human and medical genetics by the Annals of New York Academy of Sciences [68].

Expression in cell lines of mutant DARC molecules and studies of their affinity for CXCL8 and various antibodies [69-79] helped to define the coarse topology of DARC [80], i.e., a 7 transmembrane protein with four extracellular loops (named Extra Cellular Domains or ECDs) and four intracellular loops (named Intra Cellular Domains or ICDs). These experimental data coupled with sophisticated molecular modeling had allowed proposing pertinent structural models [81] able to explain the docking of DARC with CXCL-8 [82] and in the future hopefully with the malaria protein DBP.

**Principle of the in silico approach**

![Figure 1. The three steps.](image)

Transmembrane protein represents about ~25% of proteins coded by genomes. They are composed of two major classes: all-α, e.g. rhodopsin and all-β, e.g., Outer Membrane Proteins. They are the support of essential biological functions acting as receptors, transporters or channels. Moreover, 2/3 of the marketed drugs targets a transmembrane
protein and 50% specifically a GPCR [83], obtaining atomic structures becomes a major axis of research. Yet, these proteins are embedded in a lipid membrane that constitutes a very specific environment. Because the proteins are strongly destabilized when extracted from their natural medium, 3D transmembrane structures are difficult to obtain experimentally. On that account, the total number of transmembrane proteins in the Protein DataBank [84] is limited, representing ~1% of all the available structures [85]. Thus alternative approaches are required to obtain structural information. Consequently methods aiming at constructing 3D structural models are become an important research fields for understanding biological mechanisms and interactions [86].

Figure 1 shows the principle of the in silico approach adopted to build structural DARC models and to dock its ligands. It can be roughly divided into three steps. Firstly, pertinent structural models were built using comparative modeling. Then, after refinement of different models, docking of the DARC natural ligand, i.e. CXC-L8, was performed by sophisticated docking techniques using the most relevant structural models. Finally, similar docking approaches can be applied to the malarial protein DBP.
The structural models

Building a structural model of a transmembrane protein related to GPCR can be decomposed in simple tasks (cf. Figure 2 and [87]). Once the helical transmembrane regions are located with dedicated prediction software, these predicted helices are aligned with the corresponding TM helices of available GPCR structures, e.g. rhodopsin (PDB code 1F88 [88]). However few transmembrane proteins can easily fit into this idealistic pipeline. For two main reasons: (i) the sequence similarity with rhodopsin (as well with beta2-adrenergic receptor as A2A adenosine receptor) is often low and (ii) prediction of transmembrane segments requires special attention and caution.

From experimental evidence, DARC protein can be divided into three main domains: (i) one central domain encompassing the transmembrane helices and connection loops, (ii) the long N terminus region, namely ECD1 and (iii) the short C terminus region, namely ICD4.

For none of these domains, neither PSI-BLAST [89] nor FASTA [90], the most known sequence similarity search tools, retrieve protein structures related to DARC in the Protein
Data Bank Moreover DARC family has only at most 25% of identity with CXCR-2 family. Phylogenetic analyses show that DARC family is highly conserved but very different from other chemokine receptor sequences (see Figure 3, [71, 91-93]).

Figure 3. Phylogenetic analysis of DARC. Different sequences of DARC were aligned using CLUSTALW [91]. The dendogram is plotted with Njplot software [92]. Bootstrap values computed by CLUSTALW are shown at the different nodes. Sequences were extracted from UniProt/UniParc ID [94]: the major 336 aa [93] human (DUFF_HUMAN, Q16570 [95]) was used, gorilla (Gorilla gorilla), AF311914; marmoset (Callithrix jacchus), AF311915; tamarin (Saguinus oedipus), AF311916; night monkey (Aotus trivirgatus), AF311917; squirrel monkey (Samiri boliviensis), AF311918; brown capuchin (Cebus apella), AF311919; chimpanzee (Pan troglodytes), AF311920; rhesus monkey (Macaca mulatta), AF311921; baboon (Papio papio), AF303532; gibbon (Hylobates lar) AF303533 (all the last sequences are from [71]).

Surprisingly; when DARC fragments are considered instead of the complete protein, only short similar fragments mainly related to globular β-sheets are found in the PDB. These results are in conflict with most of the data available indicating that the central domain of DARC receptor is mainly composed of α-helices. Actually, alignment of DARC and rhodopsin is exactly at the level of a random alignment (12%). The prediction of transmembrane segments is also quite a complex problem. Numerous methods are at present available. They differ on the learning data used and on the kind of
learning and prediction algorithms. For instance, HMMTOP [96, 97] and TMHMM [98] are based on hidden Markov models. Yet, they give different results, even if the expected prediction rate—in regards to the literature—is supposed to be quite impressive. We have tested a dozen of these methods: DAS [99], TopPred 2 [100, 101], HMMTOP [96, 97], TMHMM 1.0 and 2.0 [98], PHDtm [102, 103], TMpred [104], SOSUI [105-107], SPLIT [108, 109], Pred-TMR 1.0 and 2.0 [110, 111], TMAP [112, 113], TSEG [114], TM-FINDER [115], UMDHMM\textsuperscript{TMHP} [116], MEMSAT [117, 118], PRODIV-TMHMM [119] and MemBrain [120]. Figure 4 summarizes the results obtained with most of them for DARC sequence [121]. Only the first helix is predicted with a large consensus while the fourth helix has a huge range of definitions, diverging sometimes by 15 residues.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Prediction of transmembrane segments of DARC. Are shown the predictions done using HMMTOP [96, 97], TMHMM 2.0 [98], TMAP [113], MEMSAT [117], SOSUI [107], TMpred [104], TSEG [114], TM-FINDER [115], Pred-TMR 2.0 [111], SPLIT [109] and DAS [99]. The final selection corresponds to the positions finally obtained after multiple tests and evaluations. The plot was done using R software [121].}
\end{figure}

So, we designed a supervised approach to obtain pertinent structural models. This
approach benefited from a crucial element, *i.e.* the availability of experimental data that guided the whole procedure. These data consisted in the measure of the CXCL8 binding rate with 40 mutants. These measures underline the potential accessibility of some residues. Without these experimental data, assessment of the resulting structural models would be extremely difficult. Data were collected from the literature and our lab [71, 72, 122]. Figure 3 describes the methodology we have used to obtain pertinent structural models of DARC.

---

**Figure 5. Building structural models of DARC.** (a) Prediction of transmembrane helices. (b) Alignment of helical regions with corresponding regions of rhodopsin structure. (c) Potential structural templates for *ECD1* and *ICD4* prediction of (d) Addition of these results to the complete alignment for comparative modeling. (e) Structural model generation and refinement of these models. (f) Accessibility computation of amino acids and known to be exposed. (g) In regards to the results, the alignment is modified. (h) At last, some models are selected.

A rough localization of the DARC transmembrane helical segments is done with different prediction algorithms presented in Figure 4 (cf. Figure 5a). The predicted α-helices of DARC are aligned to the assigned helices of the rhodopsin protein (PDB code: 1F88, chain
A [88]). Secondary structure assignment was done using DSSP software [123] (cf. Figure 5b). Similarly, predicted interconnecting loops of DARC were assigned systematically to their counterparts of rhodopsin. Potential structural homologues of ECD1 and ICD4 were searched with sophisticated approaches (Figure 5c) and the resulting propositions were added as additional templates (Figure 5d, see below).

The alignments with the sequences of the structural templates were done with Clustalw software [91] and manually modified. Hundred models were generated using Modeller software [124-126]. Each model was then refined, e.g., using SCWRL [127] and GROMACS [128, 129] (cf. Figure 5e) and the residue accessibility was computed with Naccess software [130] (cf. Figure 5f). The alignment was modified in the light of the biological data known. Hence, different alternative locations of helical regions were tested and the corresponding alignments were modified accordingly (cf. Figure 5g). More than ten iterations of steps 5e to 5g were done and finally, models corresponding at best to the biological data were selected (cf. Figure 5h).

As notified in Figure 5c, we made a specific research for the longest extra-membranous domains of the protein, namely ECD1 and ICD4. These two regions are exposed to solvent and consequently, they exhibit physico-chemical features similar to those of globular proteins. Secondary structure prediction methods, e.g. [131] predicted them as coil region with a medium confidence index. Turns prediction underlined many turns in both fragments [132]. HMMSTR, which is based on a more complex approach, predicted a short β-strand of 4-5 residues in ECD1, and a helical structure in ICD4, but the confidence index was poor [133].

Threading, ab initio and de novo methods were so required to find potential related protein fragments. The solutions were different but systematically, a high β-strand propensity was indicated in the first part of ECD1 and a helical structure at its end, close to the connection with the TM1 segment. In this research, we also used our own prediction approach
Protein Blocks

The classical description of protein structures involves two regular states, the α-helices and the β-strands and one non-regular and variable state, the coil. Nonetheless, this simple definition of secondary structures masks numerous limitations. Indeed, three states may oversimplify the description of protein structure; 50% of all residues, i.e., the coil, are not described even if it encompasses recurrent and similar local protein structures. Description of local protein structures have hence focused on the elaboration of complete sets of small prototypes or "structural alphabets" (SAs), that help to analyze local protein structures and to approximate every part of the protein backbone [134-137]. The principle of a structural alphabet firstly consists in designing a set of average recurrent local protein structures, able to approximate efficiently, every part of known structures (see Figure 6). Each residue being associated to one of these prototypes, then so, the 3D information of the protein structures can

Figure 6. Encoding of the protein structures (3D) in terms of Protein Blocks.
be translated as a series of prototypes (letters) in 1D, as the amino acid sequence.

Our structural alphabet is composed of 16 structurally averaged protein fragments that are 5 residues in length, called Protein Blocks (PBs, [138]). They have been used both to describe the 3D protein backbones [139] and to perform a local structure prediction [138, 140-143]. Our works on PBs have proven their efficiency in the description and the prediction of long fragments [140, 144-147] and short loops [148], to define a reduced amino acid alphabet dedicated for mutation design [149], to analyze protein contacts [150] or in the building of a transmembrane protein [81]. We have also used protein blocks to compare / superimpose protein structures [151-153] (see Figure 1b).

Developments made in other laboratories, using PBs, have focused on the reconstruction of globular protein structures [154], the design of peptides [155], the definition of binding site signatures [156], novel prediction methodologies [157, 158] and fragment-based local statistical potentials [157]. The features of this alphabet have been compared by Karchin et al. [159] with those of 8 other structural alphabets showing that our PB alphabet is highly informative, with the best predictive ability of those tested. It is nowadays the most widely used SA in the world.

*ECD1* prediction done with Protein Blocks method confirmed that the sequence is weakly informative. Nonetheless, it helped to distinguish 5 zones corresponding to (i) a first short region of extended structures, (ii) a second region related to extended structures following a short kink (iii) a third poorly informative region with a turn, confirmed by [132], (iv) a fourth fuzzy non-helical region (v) a fifth region predicted as mainly helical, with high confidence index, followed by a short loop.
Two selected models

Finally, two models were selected (see Figure 7). Interestingly, the final positions of transmembrane helices gave a better sequence identity with rhodopsin structure than initially computed using classical alignment tools, *i.e.* 25% of sequence identity for the helices. The two models were quite different; in the first one, $ECDI$ looked like a C letter and was far from the rest of the protein. This conformation was qualified as **“open”** (see Figures 7a and 7b). In contrast, in the second model, $ECDI$ was more compact with a position closer to the rest of the protein. The conformation of the second model was denominated **“closed”** (see Figures 7c and 7d). Both models encompassed most of the 27 amino acids supposed to be
exposed. Exceptions were related to two amino acids found buried while not involved in disulfide bridges. The first one is a Phenylalanine (F65) located in the beginning of the first helix. The side chain was stacked with other aromatic amino acids. The second one is an Aspartate (D263) not close to the surface. These discrepancies may be related to the quality of the models but also to the crude interpretation we made of the experimental data.

To explore the flexibility of the ECD loops, we performed simulated annealing simulations [161]. Firstly, it explains well the importance of D263 which interacts directly with the end of the ECD3. Secondly, when we analysed all the simulated structures with the use of Protein Blocks, we observed that some regions of ECD1 tent to be more helical and other ones to be more extended. Interestingly, these regions corresponded to those predicted to be more extended or more helical with the prediction tools.

![Figure 8. Normal Mode Analysis of DARC ‘open’ model. (a) (y-axis) Modes 7 to 9 for (x-axis) the whole sequence. (b) Visualization of the modes 7 to 9 on the ‘open’ model. The cross highlights the hinge region of ECD1. Visualization done with VMD software [162, 163].](image)

Experimental data show that ECD1 is involved in the binding with ligands. The quite different position of ECD1 between the two models led us to perform a preliminary study aiming at analyzing domain motions. This study could lead to alternative positions of ECD1 domain. For that purpose, normal mode analysis is well adapted. We use the WEBnm@ server developed by Reuter group [164], a service that performs an analysis of the first
Normal Mode with an efficient approach [165, 166]. It uses MMTK library [167] and an appropriate force field [168]. The three first lower modes, i.e., modes 7 to 9, showed clearly that ECD1, is involved in large domain motions with the largest contributions to the amplitude of motions (see Figure 8) with the largest contributions. The directions of motions indicated that ECD1 tends to come closer to the other ECDs. Interestingly, this approach highlights the distinction in three regions already observed with a first structured zone, a transition region and a last structured zone. The median region seems to play a role of hinge between the two extremities of ECD1 (see Figure 8b). However, these results even if quite interesting has to be considered with caution. Indeed, normal mode analysis strongly depends on the 3D structure. So, further analyses have to be performed on alternative models to confirm the ECD1 motion tendencies.

Figure 9. Visualization of electrostatics potential on the surface of DARC ‘open’ model. Visualization done with PyMol software [160].

Electrostatic potentials of DARC model were calculated using the finite-difference
Poissón-Boltzmann (FDPB) method [169]. Figure 9 presents the electrostatic potential mapped onto the molecular surface of the DARC model. Two distinct zones can be observed for the potential interaction zones of CXCL8. The first one is highly negative, the second one is highly positive. The residues in the core of the negative zone are the F22-E23 (corresponding to important epitope Fy6). Interestingly, electrostatics analysis performed for the CXCL8 dimer and monomer showed a complementary with this distribution: two major regions in terms of electrostatic potential can be delineated, one positive and the remainder negative, encompassing the loop 40s. Electrostatic potential computed for CCL5 (RANTES) showed less contrasted results; moreover the charge repartition was very different from CXCL8.

These analyses corroborate other studies that highlighted the importance of electrostatics in the binding of chemokines [170, 171]. For instance, the region of interaction of CXCR-1 which has been partially co-crystallized is positively charged, as the ECD1 N-terminus, and interacts with the loop 40s, negatively charged [172]. So, these results are coherent with all the available data.

Overall, these results show that the models we proposed provide some interesting clues for locating the binding site for natural ligands and the forces that could govern the interaction. However, a precise location, i.e. the ensemble of residues that participate to the interaction, necessitates more sophisticated methods such as docking methods. Docking algorithms allow scoring of each position of the ligand rolling onto the entire surface of the receptor. It is necessary to remind that, in case of membrane proteins, the membrane hides numerous sites of the protein. So, a systematic exploration of the receptor surface will waste a lot of time in non relevant locations. Most docking tools first consider the partners as rigid bodies and eventually introduce flexibility at the end of the procedure and generally only for some selected conformations of the complex. Eventually, the ligand (the smallest partner) can be considered as flexible. Flexibility is often limited to side chains repositioning. In some cases,
the polypeptide backbone is optimized by few steps of energy minimization. Thus, taking profit from the ICM software (reference) that allows introducing flexibility from the very first stages, we develop an original strategy to predict binding sites and complex conformations of DARC with natural ligands, \textit{i.e.} CXCL8.

### The docking approach

![Diagram of the docking approach](image)

**Figure 10. A divide-and-conquer approach.** The ECD1 (N terminus of DARC) and the Transmembrane Domain (TMD) are considered separately. (a) A rigid docking is performed on the TMD while (b) a flexible docking is done for ECD1. Then (c) a flexible docking is done to combine both results. (d) the most interesting results are analyzed.

As shown in the previous sections, DARC is a highly complex protein. To take into account the high flexibility of ECD1, we have designed a dedicated docking approach that combines rigid and flexible docking. First, a rigid-body docking was performed between CXCL8 and DARC without ECD1, \textit{i.e.}, Transmembrane Domain (TMD) encompassing the seven transmembrane helices, the extracellular connecting loops ECD2, ECD3, ECD4 and all the ICDs intracellular connecting loops (see Figure 10a), even if these last ones do not
participate in the binding. The presence of all the ECDs is very important. Indeed, the shape of the receptor molecular surface is dependent on their presence.

In parallel, a flexible docking was performed considering ECD1 domain as flexible rolling onto the surface of CXCL8 structures, maintained rigid (see Figure 10b). As no clear experimental evidence asserts if the interaction concerns mono- or dimeric form of CXCL8, both hypotheses were tested for rigid and flexible docking. The best results of each approach were then combined and a flexible docking is done where only ECD1 is flexible (see Figure 10c). Then the best results are selected and analyzed (see Figure 10d).

**Rigid docking**

![Rigid docking examples](image)

**Figure 11. Rigid docking examples.** (a – c) Three different examples of rigid docking results between TMD and CXCL8 (monomer form). (d - f) Three different views of a complex obtained by rigid body between TMD and CXCL8 (dimer form).
To screen the different potential interaction conformations between TMD and CXCL8, rigid docking simulations were performed. Two different tools, namely ClusPro [173, 174] and ICM [175, 176], were used. As numerous solutions are generated by each method, clusters of equivalent solutions have been done. Then different filters were applied to select the best solution. A first filter consists in selecting clusters where the ligand is located on the extracellular side, *i.e.* discarding conformations where ligand was in contact with ICDs or with the TMD. Indeed, it was not possible to introduce constraints to avoid *a priori* these positioning. Then, we applied experimental filters, *i.e.* the residues 51, 276, 129, 195 … are known to be potentially involved directly in the interaction between DARC and CXCL8 (see Figures 11a to 11d for some examples). One position of CXCL8 was in good agreement with the available biological data (see Figure 12). Moreover, in this conformation CXCL8 is able to bind glycosaminoglycans (GAG). GAGs - chemokine interaction play a key role in the aggregation of chemokines, in the gradient-generating release of the chemokines in the circulation, and their presentation to the receptors [177-179]. The modulation of such interactions may represent a therapeutic approach in inflammatory disease [180]. Around 400 different conformations were analyzed during this step. Results for CXCL8 monomer and dimer were both compatible.
Figure 12. Selected monomer CXCL8 – DARC TMD docking. Some important residues are highlighted.
Flexible docking

![Flexible docking](image)

**Figure 13. Flexible docking.** Some snapshots of the flexible search of ECD1 around the rigid structure of CXCL8 monomer are shown.

In parallel to the rigid docking, a docking was done where ECD1 was allowed to adapt its conformation when interacting with CXCL8. This flexible docking was done using the biased Monte-Carlo (MC) search of ICM [175, 176] based on a. ECD1 is firstly placed by a preliminary rigid body docking. Then, a random sampling is done in the $\phi$-$\psi$ space, *i.e.*, one couple $\phi$-$\psi$ is chosen in the favored Ramachandran regions and affected to one residue. The corresponding energy is computed, and then the Metropolis criterion is applied to keep or discard the conformation so generated. One million of random has been performed and 250,000 MC steps. Privileged conformations were clustered, leading to 50 clusters (see Figure 13). Three clusters were in accordance with the biological data, *i.e.* residues of CXCL8 and of DARC supposed to be implicated in the interaction. The contacts observed between ECD1
and CXCL8, were analyzed using Protein Blocks approach [138, 139].

**Complete assembly**

![Complete assembly](image)

**Figure 14. A good conformation.** ECD1, CXCL8 and TD are shown with different grey tones.

Spatial position of CXCL8 obtained from the rigid docking was combined with the results of flexible docking. The main task consisted in selecting *ECD1* conformations where the peptide bond between the last residue of *ECD1* and the first residue of TMD can be easily
made. The closure of the bond was obtained by a minimization with constraints. A new flexible docking (only ECD1 was flexible) was done to take into account the whole system. This simulation generates ~120,000 conformations which were grouped and filtered according to the biological data. For examples, 62 clusters of distinct conformations were selected for DARC Tyrosine 30 with CXCL8. Three conformations were associated to the lowest energy. Figure 14 shows the most representative conformation (the most frequently observed during the simulation) and that presenting the best compromise between favorable energy versus number of occurrences (see Figure 14).

Overall, even if the results are preliminary, the strategy described here yields interesting information that complement the biological data available.

**Future works**

The selected models need to be deeply and carefully analyzed and especially at the level of the amino acids involved at the interface between DARC and CXCL8. Concerning the DBP, the structure is incomplete but available [181]. This structure brings some hints [182, 183] to understand the invasion mechanism. A preliminary study performed for DARC-DBP complex, using a similar approach to that described above give remarkable results highly compatible with the main interaction hypothesis [183]. Models need to be refined using classical molecular modeling. Moreover, we have identified different pockets interesting for the interaction; further deeper analyses would help to understand their roles and their importance in the binding. Using these different results, it will be possible to propose designed mutants that can be tested first *in silico.*
in silico DARC

Acknowledgments

This work was supported by National Institute for Health and Medical Care (INSERM), University Paris 7 – Denis Diderot and National Institute for Blood Transfusion (INTS). LA benefits from a grant of the Région Ile-De-France. AdB and OB acknowledge Polonium grant 14287VM “Refinement of the studies on specificity of monoclonal antibody 2C3 directed against Duffy/DARC protein”.

Figure Legends

Figure 1. The three steps. (1) by comparative modeling, structural models are building; (2) docking approach is performed with the natural ligand structure, i.e., CXC-L8 and (3) with the Duffy Binding Protein (DBP).

Figure 2. Principles of homology modeling for a transmembrane protein.

Figure 3. Phylogenetic analysis of DARC. Different sequences of DARC were aligned using CLUSTALW [91]. The dendogram is plotted with Njplot software [92]. Bootstrap values computed by CLUSTALW are shown at the different nodes. Sequences were extracted from UniProt/UniParc ID [94]: the major 336 aa [93] human (DUFF_HUMAN, Q16570 [95]) was used, gorilla (Gorilla gorilla), AF311914; marmoset (Callithrix jacchus), AF311915; tamarin (Saguinus oedipus), AF311916; night monkey (Aotus trivirgatus), AF311917; squirrel monkey (Saimiri boliviensis), AF311918; brown capuchin (Cebus apella), AF311919; chimpanzee (Pan troglodytes), AF311920; rhesus monkey (Macaca mulatta), AF311921; baboon (Papio papio), AF303532; gibbon (Hylobates lar) AF303533 (all the last sequences are from [71]).

Figure 4. Prediction of transmembrane segments of DARC. Are shown the predictions done using HMMTOP [96, 97], TMHMM 2.0 [98], TMAP [113], MEMSAT [117], SOSUI [107], TMpred [104], TSEGG [114], TM-FINDER [115], Pred-TMR 2.0 [111], SPLIT [109] and DAS [99]. The final selection corresponds to the positions finally obtained after multiple tests and evaluations. The plot was done using R software [121].
**Figure 5. Building structural models of DARC.** (a) Prediction of transmembrane helices. (b) Alignment of helical regions with corresponding regions of rhodopsin structure. (c) Potential structural templates for \( ECD1 \) and \( ICD4 \) prediction of (d) Addition of these results to the complete alignment for comparative modeling. (e) Structural model generation and refinement of these models. (f) Accessibility computation of amino acids and known to be exposed. (g) In regards to the results, the alignment is modified. (h) At last, some models are selected.

**Figure 6. Encoding of the protein structures (3D) in terms of Protein Blocks.**

**Figure 7. The two DARC selected models.** Two views of (a-b) open form structural model and (c-d) closed structural model. Visualization done with PyMol software [160].

**Figure 8. Normal Mode Analysis of DARC ‘open’ model.** (a) (y-axis) Modes 7 to 9 for (x-axis) the whole sequence. (b) Visualization of the modes 7 to 9 on the ‘open’ model. The cross highlights the hinge region of \( ECD1 \). Visualization done with VMD software [162, 163].

**Figure 9. Visualization of electrostatics potential on the surface of DARC ‘open’ model.** Visualization done with PyMol software [160].

**Figure 10. A divide-and-conquer approach.** The \( ECD1 \) (N terminus of DARC) and the Transmembrane Domain (TMD) are considered separately. (a) A rigid docking is performed on the TMD while (b) a flexible docking is done for \( ECD1 \). Then (c) a flexible docking is done to combine both results. (d) the most interesting results are analyzed.

**Figure 11. Rigid docking examples.** (a – c) Three different examples of rigid docking results between TMD and CXCL8 (monomer form). (d - f) Three different views of a complex obtained by rigid body between TMD and CXCL8 (dimer form).

**Figure 12. Selected monomer CXCL8 – DARC TMD docking.** Some important residues are
in silico DARC

highlighted.

**Figure 13. Flexible docking.** Some snapshots of the flexible search of ECD1 around the rigid structure of CXCL8 monomer are shown.

**Figure 14. A good conformation.** ECD1, CXCL8 and TD are shown with different grey tones.

**References**

in silico DARC


[69] Le Van Kim C., Tournamille C., Kroviarski Y., Cartron J.P., Colin Y., The 1.35-kb and 7.5-kb Duffy mRNA isoforms are differently regulated in various regions of brain, differ by the length of their 5' untranslated sequence, but encode the same polypeptide, Blood 90 (1997) 2851-2853.


[99] Cserzo M., Wallin E., Simon I., von Heijne G., Elofsson A., Prediction of


[119] Viklund H., Elofsson A., Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information,


[138] de Brevern A.G., Etchebest C., Hazout S., Bayesian probabilistic approach for
[159] Karchin R., Cline M., Mandel-Gutfreund Y., Karplus K., Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry,
