ATOMDANCE: machine learning denoising and resonance analysis for functional and evolutionary comparisons of protein dynamics

Gregory A. Babbitt[1,*], Madhusudan Rajendran[1] , Miranda L. Lynch[2] , Richmond Asare-Bediako[1], Leora T. Mouli[1], Cameron J. Ryan[3], Harsh Srivastava[4], Kavya Phadke[1], Makayla L. Reed[1], Nadia Moore[1], Maureen C. Ferran[1] and Ernest P. Fokoue[5*]

[1]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA.

[2]Hauptmann Woodward Medical Research Institute, Buffalo, NY , USA.

[3]McQuaid Jesuit High School Computer Club, Rochester, NY 14618, USA.

[4]New York University, Rochester, NY 14618, USA.

[5]School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA.

*Correspondence: gabsbi@rit.edu, epfsms@rit.edu

Abstract –

Comparative methods in molecular biology and molecular evolution rely exclusively upon the analysis of DNA sequence and protein structure, both static forms of information. However, it is widely accepted that protein function results from dynamic shifts in machine-like motions induced by molecular interactions, a type of data for which comparative methods of analysis are challenged by the large fraction of protein motion created by random thermal noise induced by the surrounding solvent. Here, we introduce ATOMDANCE, a suite of statistical and kernel-based machine learning tools designed for denoising and comparing functional motion states of proteins captured in time-series from molecular dynamics simulations. ATOMDANCE employs interpretable Gaussian kernel functions to compute site-wise maximum mean discrepancy (MMD) between learned features of motion (i.e. functional dynamics) representing two protein states (e.g. bound vs. unbound, wild-type vs. mutant). ATOMDANCE derives empirical p-values identifying functional similarity/difference in dynamics at each amino acid site on the protein. ATOMDANCE also employs MMD to contextually analyze potential random amino-acid replacements thus allowing for a site-wise test of neutral vs. non-neutral evolution in the divergence of dynamic function in protein homologs. Lastly, ATOMDANCE also employs mixed-model ANOVA combined with graph network community detection to identify functional shifts in protein regions that exhibit time-coordinated dynamics or resonance of motion across sites. ATOMDANCE offers a user-friendly interface and requires as input only single structure, topolopy and trajectory files for each of the two proteins being compared. A separate interface for generating molecular dynamics simulations via open-source tools is offered as well.

Keywords – comparative method, machine learning, molecular dynamics simulation, protein evolution, protein resonance, thermal noise, protein function

Lay Audience Summary – ATOMDANCE is software designed to comprehensively simulate, calculate and compare protein motions two functional or evolutionary states while controlling for random noise. It is useful for finding amino acid sites on a given protein that are important in binding other proteins, DNA, or drugs/toxins. It can also be used to assess the effect of genetic mutation on protein motion. It also identifies regions of sites on proteins that tend to move together or 'resonate' as a whole unit.

Introduction –

Protein sequences are typically less conserved over evolutionary timescales than are protein structures (Illergård et al., 2009; Sousounis et al., 2012). This implies that the most important functional information contained in protein sequences is often manifested at the level of protein structure and the motion dynamics of their structural interactions. However, while protein specificity during molecular interactions has often been compared to a static lock and key (Fischer, 1894), all proteins and many ligands they interact with exhibit extensive soft matter physical properties that imply that protein-ligand function cannot be entirely understood through the analysis of protein-ligand structure alone (Koshland, 1958; Tripathi and Bankaitis, 2017). Often proteins are described as analogous to nanoscale-sized machines (Abendroth et al., 2015; Flechsig and Mikhailov, 2019; Strong, 2004), where non-random repetitive motions, which should be discernable from random thermal noise, are key characteristics of protein function. Thus the functional evolution of the cell must depend heavily upon how protein structures alter or shift their dynamic motions during important motor or logic gate functions required when proteins and other biological macromolecules collectively assemble their dynamic behaviors to form key metabolic or regulatory pathways (Babbitt et al., 2016).

Comparative methods of analysis are well developed for protein sequences and structures in the disciplines of phylogenetics (Cornwell and Nakagawa, 2017), molecular evolution (Suzuki, 2010) and structural biology (Kufareva and Abagyan, 2012). And they are deep rooted in experimental biology and its association with the historical development of the field of modern statistics (Parolini, 2015). In molecular evolution and comparative genomics, it is important to note that these methods are often applied in a specific site-wise manner as evolutionary changes to proteins through genetic mutation tend to act independently at individual sites over time. Site-wise analyses of root mean square deviation (RMSD) of superimposed protein chains are also very common in structural biology. However, site-wise comparative methods for application to molecular dynamics are only now beginning to be developed (Babbitt et al., 2020, 2018). Important types of site-wise dynamics comparisons taken from two different molecular dynamics simulations might include comparing the same amino acid site in two functional states (e.g. bound vs unbound to drugs, toxins, nucleic acids, or other proteins), comparing the same sites at two different temperatures in an investigation of thermostability, comparing the same amino acid site in two different evolutionary lineages or before and after a significant mutation event, or comparing the same amino acid site in two different epigenetic states (e.g. involving phosphorylation or methylation). Alternatively, important types of site-wise dynamics comparisons taken from the same molecular dynamic simulation could include comparing the dynamics of two adjacent or non-adjacent amino acid sites over time in order to ascertain how coordinated or 'resonant' they are in their dynamic behavior, either because of allostery or native contact. Being able to make site-wise determinations of similarities and differences in protein dynamics has potential application to the fields of computational pharmacology and vaccine development as they interface with the basic science of molecular evolution. Potential applications include the problems of identifying/predicting single protein sites involved in the evolution of vaccine or drug escape or the problem of whether a generic drug interacts similarly enough to a patented product across humans and model lab animals given their respective genetic and epigenetic differences in protein targets.

A major challenge with attempting to functionally analyze the dynamic trajectories of atoms in molecular dynamic simulations is caused by the large fraction of motion in the system that is simply due to thermal noise, and therefore not machine-like in its manifestation. This is especially problematic in explicit solvent based approaches, which often more accurately replicate dynamics than implicit solvent methods (Darden et al., 1993; Petersen, 1995; Wang et al., 2012). Thermal noise in explicit solvent MD simulation is caused by the random collision of water molecules and free ions with the protein chains. Past methods of comparative molecular dynamics analysis have relied upon a large amount of resampling of the atom trajectories to be able to resolve site-wise functional differences in dynamics caused by binding interactions and mutations (Babbitt et al., 2020, 2018). Despite their promise for potentially filtering random from non-random (i.e. functional) motion in MD simulation, site-wise machine learning approaches remain largely unexplored for this purpose; but see (Babbitt et al., 2020) for one approach to identifying regions of conserved dynamics.

The motion of proteins over relatively short time scales in MD simulations is largely harmonic and therefore we propose that they can be theoretically addressed with a Gaussian kernel-based approach to machine learning. Because proteins are polymers, the motions of constituent atoms are often restricted by steric constraints caused by the protein folding. This creates unequal variances in the degree of atom fluctuation (i.e. directionless magnitude of motion) across sites. This harmonic oscillation with unequal variance in atom motion in protein polymer chains implies that a machine learning approach that incorporates a Gaussian kernel function might perform well when tasked with identifying functional or evolutionary differences when comparing noisy explicit solvent MD simulations. Kernel learning also has an advantage in that it is much more interpretable than black box methods such as neural networks when applied to many physical systems (Ponte and Melko, 2017). This interpretability might be very important to biomedical researchers when deciphering the complicated protein target interactions with small molecules drugs and comparing them to natural signaling molecules they are intended to inhibit or outcompete.

Here we introduce the ATOMDANCE statistical machine learning post-processor for comparative molecular dynamics; a software suite for interpretable machine learning comparison and resonance analysis of functional protein dynamics at individual site-wise resolution. The maxDemon 4.0 program in ATOMDANCE is trained on the matrices of local atom fluctuations derived from the molecular dynamics trajectories of proteins in two functional states (e.g. bound vs. unbound or wild-type vs. mutant). The differences between dynamics at given sites are reported as maximum mean discrepancy (MMD) in the reproducing kernel Hilbert space (RKHS). The site-wise average differences and Kullback-Leibler (KL) divergences are also reported using the DROIDS 5.0 method previously developed by our lab group (Babbitt et al., 2020, 2018). Hypothesis tests for significance of functional dynamic changes reported via MMD are also provided using a bootstrapping approach. For a test of neutral evolution on protein dynamics, the MMD of observed amino acid replacements is compared to a neutral model of the MMD of random pairs of differing amino acids at different sites on the two protein simulations being compared. This allows for the identification of potential sites where natural selection has either functionally conserved or adaptively altered the local molecular dynamics of the protein. The ChoreoGraph 2.0 program in ATOMDANCE offers a site-wise mixed-model ANOVA and graph network community analysis of time interaction (i.e. resonance) for the identification of regions of amino acid sites with high of coordination in their respective dynamic states. ATOMDANCE is intuitive and user-friendly, providing a simple graphical user interface (GUI) that only requires structure, topology, and trajectory files (.pdb, .prmtop, .nc) for the two molecular dynamics simulations being compared. It is entirely python-based and outside of this it only requires UCSF ChimeraX for molecular visualization and the popular cpptraj library for resampling calculations (Goddard et al., 2018; Pettersen et al., 2021; Roe and Cheatham, 2013). ATOMDANCE is supplemented with an optional GUI for generating simulations via open-source tools (i.e. openMM and AmberTools) (Case et al., 2005a; Eastman et al., 2017). However it can also potentially be used with files generated using NAMD (qwikMD), CHARMM, or the licensed version of Amber (Case et al., 2005a; Jo et al., 2008; Phillips et al., 2020; Ribeiro et al., 2016). We provide an optional GUI for generating multiframe PDB file movies using UCSF ChimeraX (Pettersen et al., 2021) where the motions of the protein system are colored and augmented in accordance to the MMD.

In summary, the ATOMDANCE software suite provides researchers with a computational platform for supplementing comparative sequence/structure analyses with important information about the functional motions of proteins undergoing complex interactions with DNA/RNA, drugs, toxins, natural ligands, or other proteins. It can also address how genetic mutation and/or molecular evolution has altered these functional motions as well as how interactions in motions across sites are coordinated over time and how these pattern of site resonance might shift during mutation and or functional interactions with other small molecules or other large biological macromolecules in the cell.

Methods –

PDB structure and model preparation

Protein structures of ubiquitin, TATA binding protein (TBP), BRAF kinase bound to the drug sorafenib, HIV-1 main protease dimer bound to paptide, and SARS-CoV-2 receptor binding domain bound to angiotensin converting enzyme 2 (ACE2) were obtained from the Protein Data Bank (PDB). Summary of the different PDB structures used for the MD simulation runs are listed in Table 1. After downloading the structures from the PDB database, any crystallographic reflections, ions, and other solvents used in the crystallization process were removed. Any missing loop structures in the protein structures were inferred using the MODELLER homology modelling server in UCSF Chimera. pdb4amber (AmberTools20) was employed to add hydrogen atoms (i.e. reduce the structure) and remove crystallographic waters.

Molecular dynamic simulation protocols

For each molecular dynamic comparison (monomer vs. dimer, wildtype vs. mutant protease; protease bound and unbound to drug), accelerated molecular dynamic (MD) simulations were performed (Case et al., 2005b). MD simulation protocol was followed as previously described, with slight modifications(Babbitt et al., 2022a, 2022b, 2022c, 2020, 2020, 2018; Rajendran et al., 2022; Rajendran and Babbitt, 2022). In brief, for each MD comparison, large replicate sets of accelerated MD simulation were prepared and then conducted using the particle mesh Ewald method implemented on A100 and V100 NVIDIA graphical processor units by pmemd.cuda running Amber20 (Case et al., 2005a; Darden et al., 1993; Ewald, 1921; Pierce et al., 2012; Salomon-Ferrer et al., 2013) and/or OpenMM (Eastman et al., 2017). The MD simulations were done on a high performance computing workstation mounting dual Nvidia 2080Ti graphics processor units. All comparative MD analysis via our ATOMDANCE was based upon 50 randomly resampled windows collected on of 10ns of accelerated MD in each comparative state, e.g., monomer vs. dimer, wildtype vs. mutant, protease bound to drug vs. protease unbound to drug). Explicitly solvated protein systems were first prepared using teLeap (AmberTools 20), using ff14SSB protein force field, in conjunction with modified GAFF2 small molecule force field (Maier et al., 2015; Wang et al., 2004). Solvation was generated using the Tip3p water model in a 12nm octahedral water box. Charge neutralization was performed using Na+ and Cl- ions using the AmberTools22 teLeap program. Force field modifications for the small molecule ligands were generated using scaled quantum mechanical optimization via the sqm version 17 program in antechamber/AmberTools22 (Walker et al., 2008). For each MD comparison, an energy minimization was first performed, then heated to 300K for 300 pico seconds, followed by 10 ns of equilibration, and then finally a replicate set of 100 MD production run was created for each comparative state. Each MD production run was simulation for 1 ns of time. All simulations were regulated using the Anderson thermostat at 300k and 1atm (Andersen, 1980). Root mean square atom fluctuations were conducted in CPPTRAJ using the atomicfluct command (Roe and Cheatham, 2013). All molecular color-mapping of our results were conducted in UCSF ChimeraX (Goddard et al., 2018; Pettersen et al., 2021). Any x-ray crystal protein structures requiring missing loop refinement were corrected using MODELLER prior to preparation for molecular dynamics simulations (Sali and Blundell, 1993).

ATOMDANCE.py is a PyQt5 GUI designed for post-processing comparative molecular dynamics and delivering information about important protein site differences between the dynamics of proteins in two different functional states. It also can be used to investigate potential site-wise evolutionary changes in protein dynamics and to investigate where sites share coordinated dynamics states as well. After randomly subsampling the atom trajectory files and calculating amino acid site atom fluctuations and atom correlations using the atomicfluct and atomiccorr functions from the cpptraj library, ATOMDANCE.py runs 4 types of analyses listed below.

DROIDS 5.0 – protein site-wise divergence metrics for pair-wise comparison of protein backbone atom fluctuations across functional protein states (e.g. bound vs. unbound or wildtype vs. mutant)

This option is an acronym for Detecting Relative Outlier Impacts in Dynamics Simulations and calculates both the average differences and KL divergences in the atom fluctuation at every protein site. Fluctuations are averaged over the four protein backbone atoms for each amino acid (i.e. N, Cα, C, and O). Significant differences in dynamics of the two protein states are determined by a two sample Kolmogorov-Smirnov test corrected for the number of sites in the protein corrected for the false discovery rate (i.e. Benjamini-Hochberg method) caused by the total number of sites on the protein. This method is described and published previously in DROIDS v1.0-4.0 (Babbitt et al., 2020, 2018). The only difference in v5.0 is that the subsampling is taken from random window positions along a single long MD production run, rather than multiple short MD production runs.

maxDemon 4.0 – protein site-wise kernel learning for pair-wise comparison of protein backbone atom fluctuations and/or atom correlations across functional protein states (e.g. bound vs. unbound or wildtype vs. mutant)

This analysis option uses site-wise training of Gaussian processes machine learners with tuned radial basis kernel functions in order to specify a maximum mean discrepancy (MMD) in reproducing kernel Hilbert space (RKHS) that describes the distance in learned features between the two protein dynamic states at all given sites on the protein.

Thus the kernel function describing the mapping of the data points x_i and x_j being compared is

$$k(x_i\, , x_j\,) = exp\left(\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)$$

And the empirical estimations of MMD, or distance between feature means is given by

$$MMD^2(X,Y) = \frac{1}{m(m-1)}\sum_i \sum_{j\neq i} k(x_i, x_j) - 2\frac{1}{m(m-1)}\sum_i \sum_j k(x_i, y_i) + \frac{1}{m(m-1)}\sum_i \sum_{j\neq i} k(y_i, y_j)$$

where x’s are the data points we have and y’s are generated examples evaluated on the kernel.

The learners in ATOMDANCE can be trained using a local atom fluctuation feature vector comprised of fluctuations from sites -2, -1, 0, 1, 2 positions on the protein chain relative to the site being analyzed. The observed site-wise MMD values are further subjected to hypothesis testing using a bootstrap derived empirical p-value whereby the observed MMD values between the functional dynamic states at any given site are compared to 500 bootstrapped MMD values for that site when derived from resampling the dynamics in the same dynamic state. A graphical overview of this analysis is shown in Figure 1. A key concept here when comparing this output to the site-wise KL divergence metrics generated by DROIDS 5.0 is that because the learner cannot optimize on random differences in atom fluctuation caused by thermal noise it acts as a noise filter, thus eliminating motion dampening that is not directly due to non-random differences in atom fluctuation between the sites being compared (i.e. functional aspects of molecular interactions directly involved in the binding interaction)

maxDemon 4.0 – protein site-wise kernel learning for detecting natural selection acting upon protein dynamics via neutral modeling of amino acid replacements (e.g. human vs. another species ortholog or another human paralog)

This analysis option is only appropriate when comparing two homologous proteins in two states of molecular evolution, whereby mutations have accrued over time and the user would like to determine whether the dynamics at a given site of amino acid replacement has likely been functionally conserved, evolved neutrally or evolved adaptively (i.e. under purifying, neutral or adaptive evolution). Comparisons of dynamics between the same protein in two different species (i.e. orthologs) or two related proteins in the same species (i.e. paralogs) are both enabled through this method of analysis. In this case, the MMD in dynamics between each site of amino acid replacement between the homologous proteins is compared to a model of neutral evolution represented by a distribution of MMD taken from the dynamics of 500 random pairs of dissimilar amino acids at different sites on the homologous proteins. If the MMD for an observed amino acid replacement is in the lower or upper extremities of the distribution of neutral MMD (two-tailed level of significance = 0.05), then natural selection acting upon the dynamics can likely be inferred.

ChoreoGraph 2.0 – identification of protein site communities with coordinated dynamic states mixed-effects model ANOVA and Louvain community detection

This analysis option examines the reference and query dynamic state simulations of the proteins compared above and produces (A) site-wise heatmaps and community level graph networks identifying groups of amino acid sites where atom fluctuation values are resonating over time in a coordinated fashion and (B) site-wise heatmaps and community level graph networks identifying groups of amino acid sites where overall atom fluctuation values are not significantly different from each other (i.e. potentially in contact). In each dynamic state simulation every site *i* on the protein is compared to every site *j* using a mixed effects model ANOVA where atom fluctuation represents a fixed effect (α) in the model and a time sample represents a random effect (β) in the model. Thus the general linear model becomes

$$Y_{st} = \mu + \alpha_s + \beta_t + \alpha\beta_{st} + \varepsilon_{st}$$

where *s* represents the site class (*i* or *j*) and *t* represents the random time sampling group collected by the cpptraj program (see subsampling step in Figure 1).

For the resonance analysis, the p-value of interaction between atom fluctuation levels between site *i* and site *j* and the time subsamples in the MD simulation (i.e. $\alpha\beta_{st}$) indicates the significance of an interaction of atom fluctuation between the two different sites over time (i.e. a coordinated physical resonance in motion). These p-values are shown as a heatmap. In the second step of the analysis intended to define communities of resonating regions of protein dynamics, the interaction p-values for all site *i* to site *j* comparisons are represented by a graph network where (k_n), the degree of node n is

$$k_n = A_{nm} \sum m$$

where *n* and *m* are the interaction p values for sites *i* and *j,* and A_{nm} is the adjacency matrix connecting nodes *n* and *m*. The Louvain community detection algorithm (Blondel et al., 2008) iterates a two step process of modularity optimization followed by community aggregation until community identities of all nodes are stable. It is implemented in our code by the python package networkx (Hagberg et al., n.d.).

For the analysis of potential adjacent and non-adjacent contacts, the same procedure as used above is repeated with the exception that the p-values no longer represent the interaction terms of the ANOVA, but now represent the fixed effect of the model. So in this case the resulting heatmap and network communities represent sites whose overall atom fluctuations are NOT significantly different from each other. In general, these networks are highly fragmented (i.e. often pairs of sites) and a trace of the non-significant p-values in the heatmap can easily be used to assess whether sites *i* and *j* are adjacent or non-adjacent to each other.

ATOMDANCE is available at GitHub/GitHub page

https://github.com/gbabbitt/ATOMDANCE-comparative-protein-dynamics

https://gbabbitt.github.io/ATOMDANCE-comparative-protein-dynamics/

Examples presented in this manuscript were generated from structure, topology, and trajectory files deposited here https://zenodo.org/record/7679282#.Y_wIK9LMJ9A

DOI 10.5281/zenodo.7679282

makeMovie.py

A supplemental python GUI program for making molecular dynamics movies that are weighted by the normalized MMD in atom fluctuation between two functional states. The program first creates a multi-frame PDB file representing the true dynamics of the protein system, then it creates a multi-frame PDB file where the noise in the trajectories is dampened or amplified according to MMD. This creates a purely visual effect in a color-mapped movie of protein motion that demonstrates what the MMD filter captures. We have demonstrated this in examples of both dampening of atom motion during TATA binding protein interaction with DNA and with amplification of motion in the activation loop of BRAF kinase during cancer drug binding. https://people.rit.edu/gabsbi/img/videos/MMDmovie.mp4

MDgui.py

We also provide a full python GUI for running MD simulations using open source AmberTools and openMM. The user can generate MD trajectory and topology files using any software they prefer. Other options include Amber (licensed), NAMD/QwikMD (free), CHARMM (licensed), or openMM (free). The cpptraj software available on GitHub or in AmberTools can be used to convert common trajectory file formats to the binary format (.nc) used by ATOMDANCE. We also offer a useful python+perl-based GUI for licensed versions of Amber available here. https://gbabbitt.github.io/amberMDgui/

Note on the naming of things:
DROIDS – acronym for Detecting Relative Outlier Impacts in Dynamics Simulations
maxDemon – abbreviated from Maxwell's Demon, a 19th century thought experiment connecting the concepts of information and entropy in thermodynamics involving a mythical demon watching/assessing the motion of every atom in a system.
ChoreoGraph – evokes a notion of when motions of atoms at amino acids site 'move together' in a coordinated manner, in much the same way dancers may move together in choreography.
ATOMDANCE – an homage to a song composition by Icelandic singer Bjork Guomundsdottir from her 2015 album Vulnicura (One Little Indian Records)

Results –

The first example of comparative protein dynamics analyses conducted with ATOMDANCE investigated the functional effect of DNA binding to TATA binding protein (TBP; PDB 1cdw) by the site-wise comparison of atom fluctuation of TBP in both its DNA bound and unbound state. Figure 2A-B shows both color-mapped protein surface and site-wise plot of the KL divergence in fluctuation (i.e. DROIDS 5.0). This protein binds quite strongly as is evidence by a general dampening of fluctuation across the protein (in blue). The comparison of machine learning derived MMD (i.e. maxDemon 4.0; Figure 2C-D) clearly captures the key sites of the functional interaction; two loops of the protein that interact directly with the major groove of the DNA (in blue). Supplemental Figures 1-2 show alternative plots generated by ATOMDANCE showing site-wise atom fluctuation profiles, amino acid types, and bootstrapped empirical p-values. These color mapped surfaces and plots indicate that the overall correlations of motions of amino acid sites are quite distinct across the proteins axis of symmetry, with the structures supporting the left and right side binding loops being distinct in their correlation patterns. The second example of the application of MMD captures the functional amplification of atom fluctuation by the activation loop (shown in red) of BRAF kinase upon the binding of the drug sorafenib in the ATP binding pocket of the kinase domain (PDB 1uwh; Figure 3). In this example the interaction of ATP or ATP-competitive drugs like sorafenib clearly amplify the motion in the activation loop. While sorafenib binds the site stronger than ATP, thus interrupting the MAPK pathway triggering cell proliferation in tumorogenesis, this amplification of the activation loop by the drug may be functionally related to the hyperactivation of MAPK in surrounding normal cells, leading to cancer recurrence (Babbitt et al., 2022c). A third example also demonstrates the utility of the maxDemon 4.0 MMD to investigate the protein-protein interaction between the viral SARS-CoV-2 receptor binding domain (RBD) and its human protein target angiotensin converting enzyme (ACE2)(PDB 6m17). The functional response of the RBD in Figure 4A-B captures its known binding region domain (in blue) that interacts with the N-terminal helices of ACE2. Figure 4C-D shows the dampening (in blue) at key sites of ACE2 that are recognized by the viral RBD. These include two sites on the N-terminal helices of ACE2 including two well documented additional sites identified at Q325 and K353 identified in previous studies of functional dynamics and evolution (Rajendran et al., 2022; Rajendran and Babbitt, 2022; Rynkiewicz et al., 2021).

To demonstrate the utility of MMD in a comparative evolutionary analysis of human vs. bacteria TBP (Figure 5), maxDemon 4.0 derived a neutral model distribution of MMD in dynamics for randomly selected pairs of differing amino acid sites on the human and bacterial orthologs (Figure 5A) with the tails indicating non-neutral evolution colored red for functionally conserved dynamics and green for adaptively altered dynamics. The two TBP ortholog structures are nearly identical (Figure 5B), and yet (Figure 5C) two regions of altered dynamics (i.e. high MMD) appear related amino acid replacements that have shifted the protein dynamics related to the loop binding regions highlighted earlier in Figure 2C-D. Most of the rest of the majority of the amino acid replacements (red bars) have occurred under the selective pressure to functionally conserve the TBP dynamics keeping the MMD low between the two orthologs.

The last ATOMDANCE example demonstrates the utility of the mixed effects model ANOVA in ChoreoGraph 2.0 for identifying regions with coordinated or time resonant dynamics (Figure 6) and for detecting similarities in dynamics driven by potential amino acid contacts (Figure 7). Here we analyzed the unbound and sorafenib drug-bound dynamics of the BRAF kinase domain and demonstrate a profound shift in regions of resonant motions that develops when binding occurs. The main effect of this shift served to connect the ATP binding pocket targeted by the drug to the activation loop of kinase domain whose motion is amplified upon binding (Figure 3). As expected, this shift in resonance does not very much disrupt the pattern of atom fluctuation similarity cause by the structure itself (Figure 7). We previously investigated the role of these dynamics in the hyperactivation of the MAPK pathway in many cancers (Babbitt et al., 2022).

Discussion -

Molecular dynamics simulation is a powerful tool for estimating physicochemical properties of systems in modern protein science. However, its utility has been limited by the lack statistically sound methods that can allow comparative functional and evolutionary analyses of protein dynamics. Unlike protein sequence and structural data, both static forms of data, capturing protein motion via molecular dynamics simulations creates a large component of variation that is induced by solvent-induced random thermal noise, subsequently creating a dataset in which non-random functional motions of proteins are obscured by this noise. We have described ATOMDANCE, a software suite for comparative protein dynamics that includes a powerful and interpretable kernel-based machine learning post-processor that allows users to mitigate the effects of noise and identify functional and evolutionary differences in molecular dynamics at individual sites on proteins. While site-wise differences and divergence metrics can capture meaningful differences in protein dynamics related to binding function, they can have a sometimes have difficult time identifying functionally important binding sites incorporating non-random machine like motions from sites moving more randomly but still altered by large overall shifts in thermodynamics affecting large parts of the protein. The TATA binding protein used in our validations, is a perfect example of such a protein, as it utilizes two functional binding recognition loops, but nevertheless binds DNA so strongly so as to greatly deform the rigid DNA molecule and dampen atom fluctuation across nearly the whole of the TATA binding protein. While divergence metrics in DROIDS 5.0 capture this overall dampening at nearly all protein sites very well, our kernel learner in maxDemon 4.0 clearly identifies the functional binding sites themselves hidden within the thermal noise. We have demonstrated the utility of ATOMDANCE for investigating a variety of functional aspects of protein systems including DNA-binding, drug/ligand induced shifts in dynamics, protein-protein interactions in infectious disease, as well as the functional evolutionary divergence/convergence of human-bacterial protein orthologs. ATOMDANCE also offers traditional comparative metrics applied to molecular dynamics as well as a novel kernel-based approach to identifying regions across the protein where changes or shifts in dynamics over time are coordinated across many sites. ATOMDANCE is entirely python-based with an easy to use graphical interface with seamless interaction with the open-source cpptraj MD analysis library and the modern UCSF ChimeraX molecular visualization software.

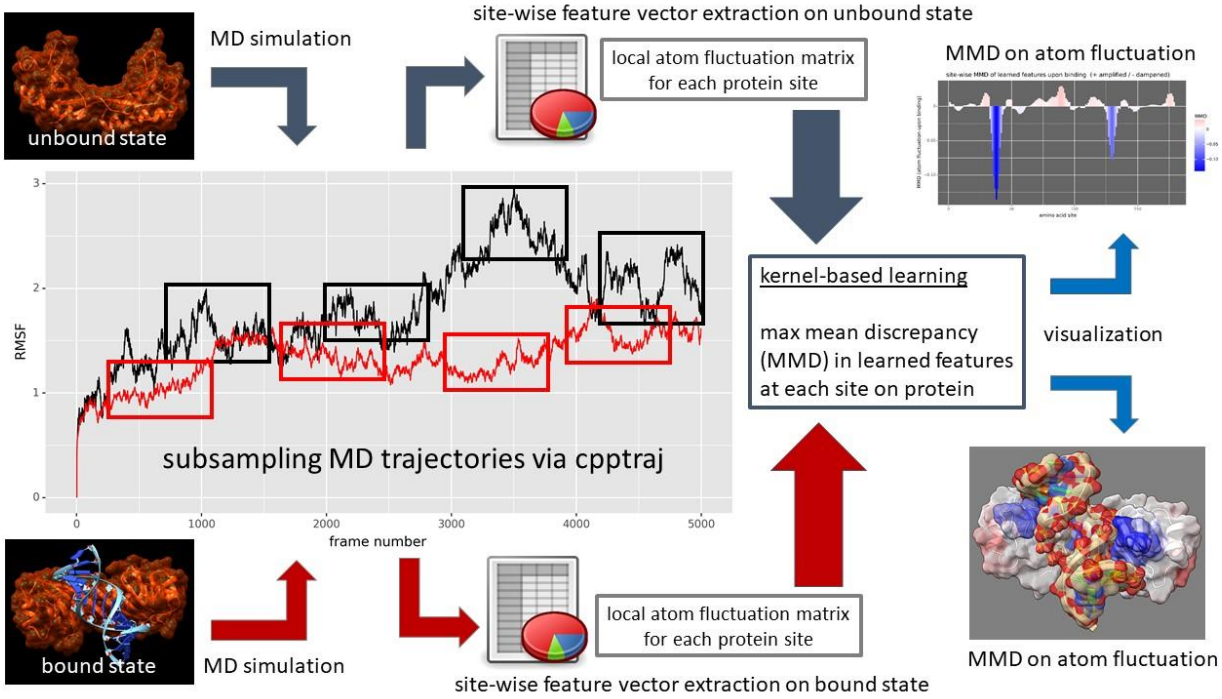
unbound state
MD simulation
site-wise feature vector extraction on unbound state
local atom fluctuation matrix for each protein site
MMD on atom fluctuation
site-wise MMD of learned features upon binding (+ amplified / - dampened)
MMD (atom fluctuation upon binding)
0
-0.05
-0.10
amino acid site
MMD
0
-0.05
-0.10
3
2
1
0
RMSF
0
1000
2000
3000
4000
5000
frame number
subsampling MD trajectories via cpptraj
kernel-based learning
max mean discrepancy (MMD) in learned features at each site on protein
visualization
bound state
MD simulation
local atom fluctuation matrix for each protein site
site-wise feature vector extraction on bound state
MMD on atom fluctuation

Figure 1 – Overview of the ATOMDANCE statistical machine learning post-processor for comparative protein dynamics. First two molecular dynamics simulations representing the functional end states are conducted (e.g. drug/DNA/protein bound vs. unbound). The .pdb structure, .prmtop topology, and .nc trajectory files for both states are input to the software and the trajectories are repeated subsampled according to user specification using cpptraj. Site-wise local atom fluctuation matrices are constructed from the subsampling and used to train a Gaussian process kernel (radial basis function). At each site on the protein, the maximum mean discrepancy (MMD) in reproducing kernel Hilbert space is calculated, representing the distance between the learned features in the transformed data space that best captures the functional difference in protein dynamics at the given site. The MMD is signed negative if atom motion is dampened or positive if it is amplified. This MMD is visualized in a variety of plots and can be color-mapped to the .pdb structure file in UCSF ChimeraX (blue indicating regions of dampened motion due to binding interaction).

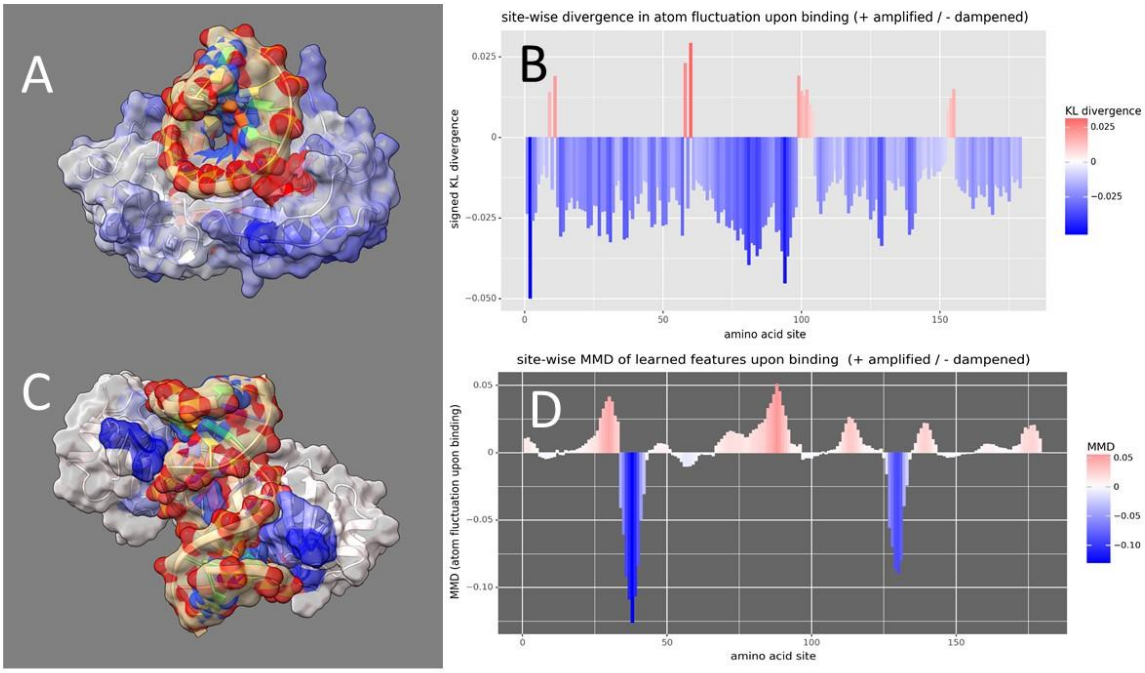

Figure 2 – Site-wise divergence metrics and signed maximum mean discrepancy (MMD) in local atom fluctuation when comparing molecular dynamics simulations of DNA-bound versus unbound TATA binding protein (PDB: 1cdw). (A-B) Signed symmetric Kullback-Leibler (KL) divergence in atom fluctuation indicating sites where motion is dampened during DNA binding (blue) and where motion is amplified (red) demonstrates generalized motion dampening across the whole protein. (C-D) Kernel-based learning applied to local atom fluctuation (i.e. signed MMD) much better isolates the loop regions of the protein that functionally interact with the DNA major groove.

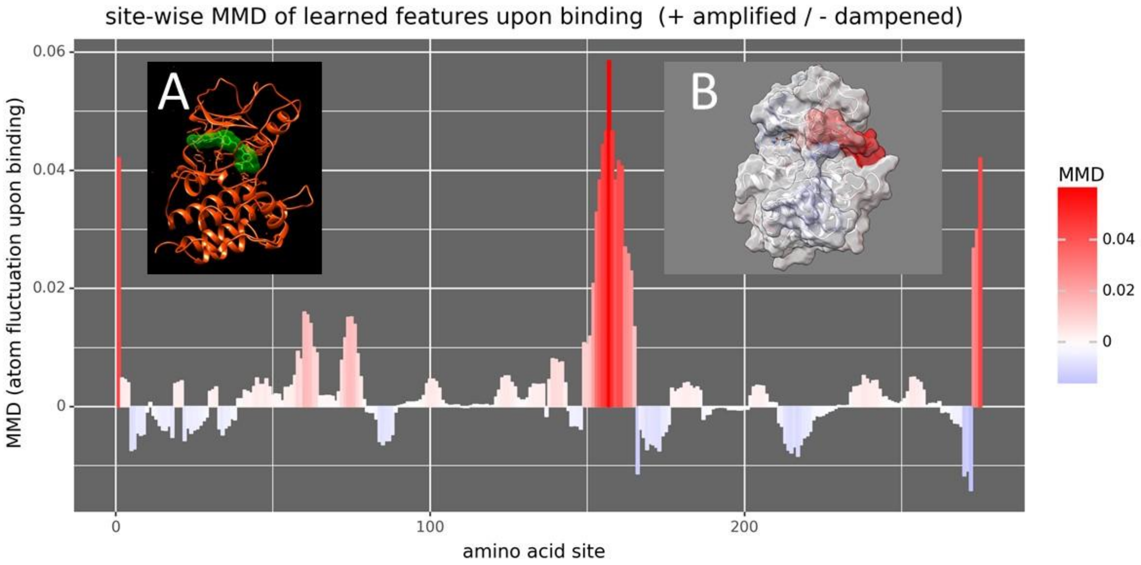

Figure 3 – Site-wise signed maximum mean discrepancy (MMD) in local atom fluctuation when comparing molecular dynamics simulations of sorafenib drug-bound versus unbound B-Raf kinase protein (PDB: 1uwh) where motion is amplified (red) or dampened (blue) during drug binding. (Kernel-based learning applied to local atom fluctuation (i.e. signed MMD) caputes the main activation loop region of the protein that responds when sorafenib (or ATP) occupies the ATP binding pocket. Insets show the (A) drug binding position in the kinase domain and (B) the location of the activation loop region on the kinase domain.

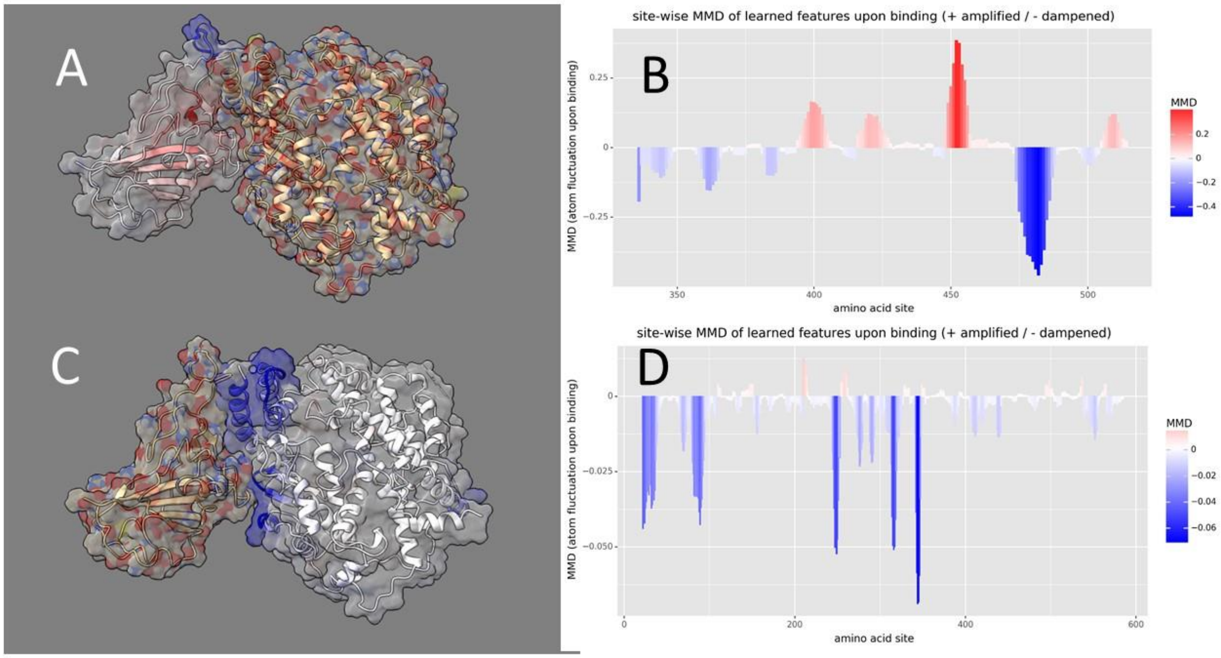

Figure 4 – Site-wise signed maximum mean discrepancy (MMD) in local atom fluctuation when comparing molecular dynamics simulations of SARS-CoV-2 receptor binding domain (RBD) in complex with angiotensin converting enzyme 2 (ACE2) versus (A-B) unbound RBD and (C-D) unbound ACE2 (PDB: 6m17). Blue regions indicate key amino acid sites on the SARS-CoV-2 and ACE2 proteins that are involved in functional viral interaction. Red regions correspond to amplified motion during viral binding likely caused by transfer of thermal energy from the tightly bound regions of the viral-host protein complex.

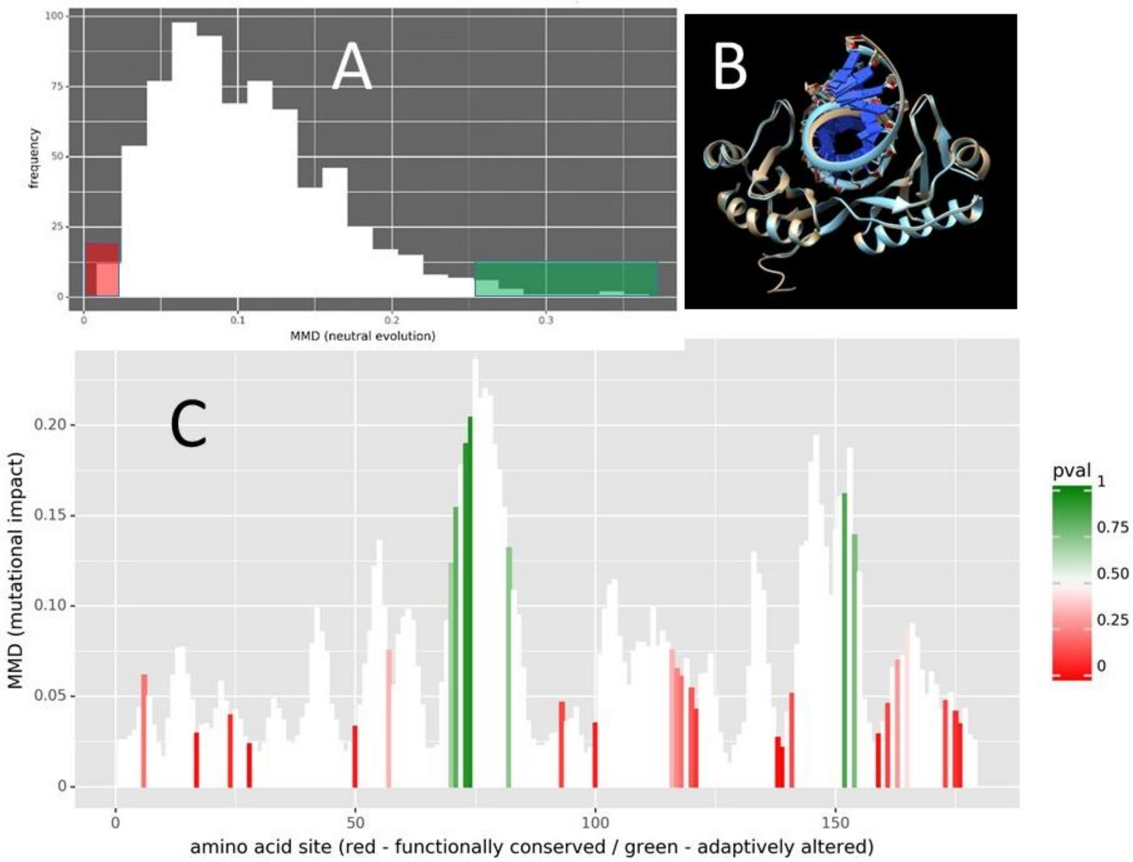

Figure 5 – Site-wise unsigned maximum mean discrepancy MMD in local atom fluctuation and atom correlation comparing DNA-bound models of bacterial and human orthologs of TATA binding protein (PDB: 1qna and PDB 1cdw resp.). As a test of neutral evolution, the MMD between dynamics on randomly chosen differing amino acid sites between the orthologs is used to generate (A) an expected distribution of MMD for the effects of random amino acid replacement on molecular dynamics. The tails of the distribution are used to identify MMD values indicative of functionally conserved dynamics (red) or adaptively altered dynamics (green). (B) The superimposition of the two structures shows that the protein has maintained near perfect structural similarity since the divergence of common ancestor between bacteria and humans despite many amino acid replacements over time. (C) The MMD profile of the dynamic differences between orthologs is the background (in white) for the bootstrap analyses of MMD for the existing amino acid replacements (in color). Red indicates dynamic changes that are significantly smaller than expected under the neutral model (i.e. functionally conserved) while green indactes dynamic changes that are significantly larger (i.e. adaptively altered).

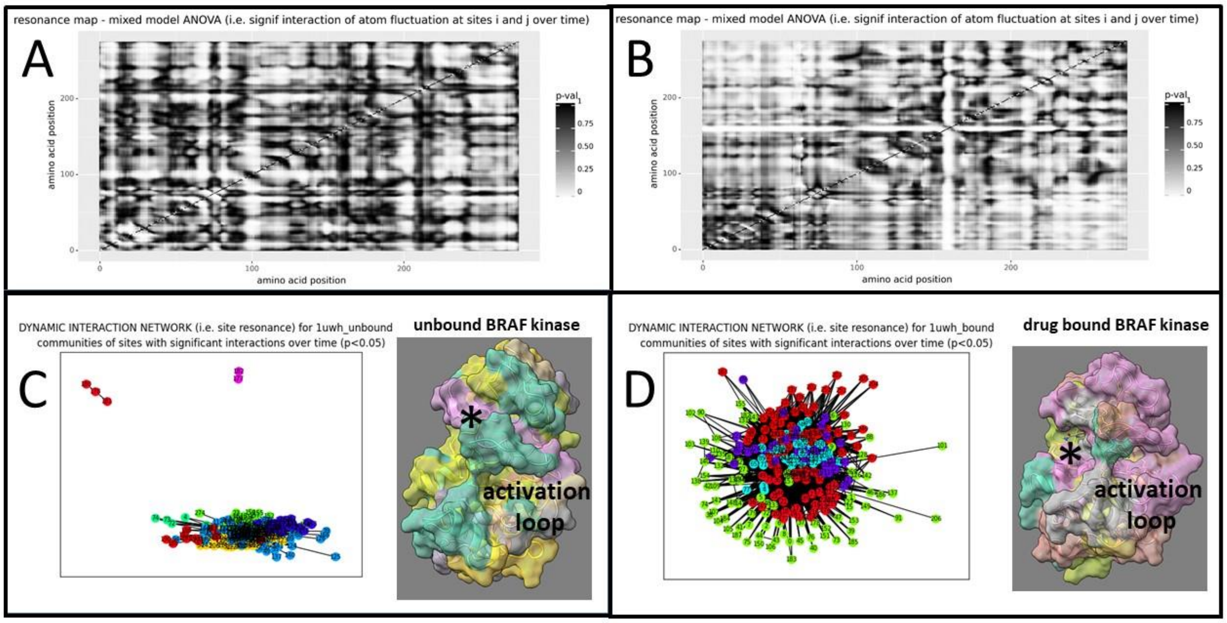

Figure 6 – Resonance analysis heat maps and community detection (ChoreoGraph 2.0) indicating regions of coordinated protein dynamics over time. The heat maps of the interaction term p-values for all pair-wise comparison of sites I to sites j on the (A) unbound and (B) drug sorafenib-bound BRAF kinase protein (PDB: 1uwh) are shown. Interaction P- values for atom fluctuation across sites over time (i.e. resonance in motions) are derived from pair-wise mixed effects model ANOVAs where atom fluctuation is the dependent variable and sites i vs site j is the fixed effect and time samples are the random effect). Resonance patterns across sites are indicated by significant p-value (white). Regions of coordinated motion derived from Louvain community detection applied to graph network analysis are shown for (C) unbound and (D) sorafenib-bound BRAF kinase. Resonance regions (i.e. communities of sites with significant time interactions) are similarly color mapped to the surface of the protein. Regions that fail to form resonance are colored light turquoise green. The ATP binding pocket that is targeted by the drug is shown with an asterisk and the activation loop whose motion is amplified upon binding is also labeled.   Note that upon binding (D) a very large resonance community/ region occurs that connects the binding site to the activation loop (pink).

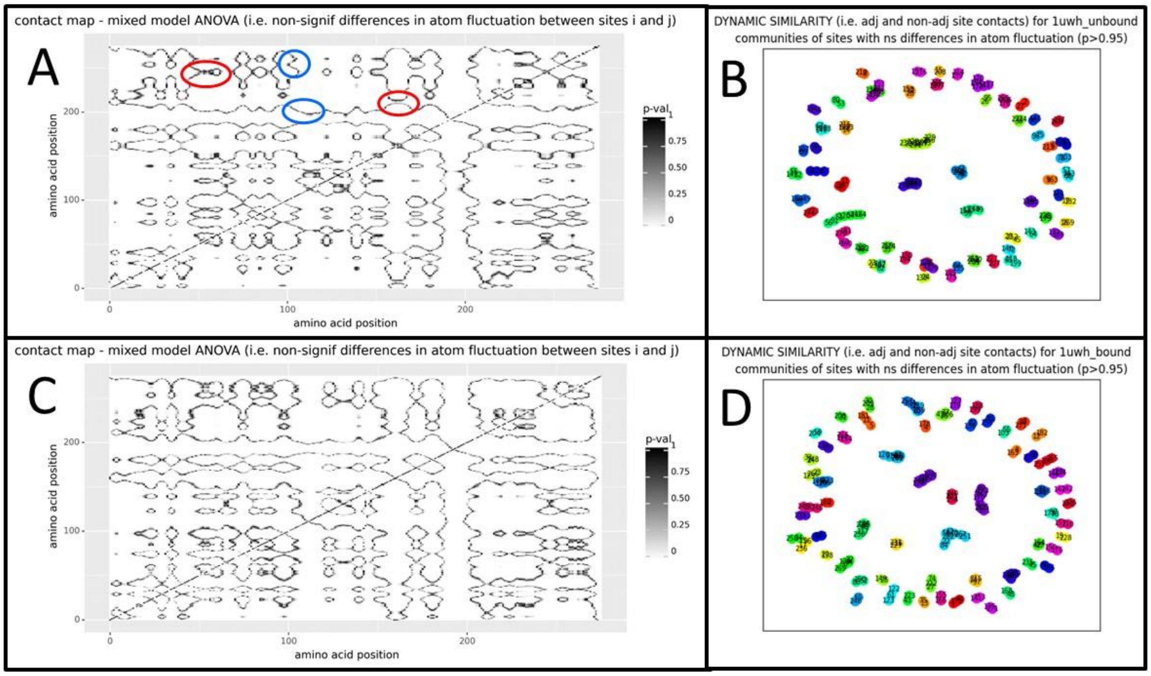

Figure 7 - Potential contact analysis heat maps and community detection (ChoreoGraph 2.0) indicating regions of overall similarity of protein dynamics (regardless of time). The heat maps of the fixed effect term p-values for all pair-wise comparison of sites I to sites j on the (A) unbound and (C) drug sorafenib-bound BRAF kinase protein (PDB: 1uwh) are shown. Non-significant p-values (black) indicate overall similarity of atom fluctuation between adjacent or non-adjacent sites. Some examples of patterns of similarity caused by adjacent sites (circle blue) and non-adjacent contacts (circled red) are shown in (A). The latter are identified where the non-significant p-value trace on the map doubles back and connects distant sites. Regions of overall similar motion derived from Louvain community detection applied to graph network analysis are shown for (B) unbound and (D) sorafenib-bound BRAF kinase. Similar regions (i.e. communities of sites with non-significant differences in overall atom fluctuation) tend to form separate communities, often only pairs or triplets of sites under the community detection algorithm.

Supplemental File – video overview with dynamics of DNA-bound TATA binding protein and sorafenib drug-bound B-Raf kinase domain weighted in accordance with maximum mean discrepancy in atom fluctuation. https://people.rit.edu/gabsbi/img/videos/MMDmovie.mp4

References

Abendroth, J.M., Bushuyev, O.S., Weiss, P.S., Barrett, C.J., 2015. Controlling Motion at the Nanoscale: Rise of the Molecular Machines. ACS Nano 9, 7746–7768. https://doi.org/10.1021/acsnano.5b03367

Andersen, H.C., 1980. Molecular dynamics simulations at constant pressure and/or temperature. J. Chem. Phys. 72, 2384–2393. https://doi.org/10.1063/1.439486

Babbitt, G.A., Coppola, E.E., Alawad, M.A., Hudson, A.O., 2016. Can all heritable biology really be reduced to a single dimension? Gene 578, 162–168. https://doi.org/10.1016/j.gene.2015.12.043

Babbitt, G.A., Fokoue, E.P., Evans, J.R., Diller, K.I., Adams, L.E., 2020. DROIDS 3.0—Detecting Genetic and Drug Class Variant Impact on Conserved Protein Binding Dynamics. Biophys. J. 118, 541–551. https://doi.org/10.1016/j.bpj.2019.12.008

Babbitt, G.A., Fokoue, E.P., Srivastava, H.R., Callahan, B., Rajendran, M., 2022a. Statistical machine learning for comparative protein dynamics with the DROIDS/maxDemon software pipeline. STAR Protoc. 3, 101194. https://doi.org/10.1016/j.xpro.2022.101194

Babbitt, G.A., Lynch, M.L., McCoy, M., Fokoue, E.P., Hudson, A.O., 2022b. Function and evolution of B-Raf loop dynamics relevant to cancer recurrence under drug inhibition. J. Biomol. Struct. Dyn. 40, 468–483. https://doi.org/10.1080/07391102.2020.1815578

Babbitt, G.A., Lynch, M.L., McCoy, M., Fokoue, E.P., Hudson, A.O., 2022c. Function and evolution of B-Raf loop dynamics relevant to cancer recurrence under drug inhibition. J. Biomol. Struct. Dyn. 40, 468–483. https://doi.org/10.1080/07391102.2020.1815578

Babbitt, G.A., Mortensen, J.S., Coppola, E.E., Adams, L.E., Liao, J.K., 2018. DROIDS 1.20: A GUI-Based Pipeline for GPU-Accelerated Comparative Protein Dynamics. Biophys. J. 114, 1009–1017. https://doi.org/10.1016/j.bpj.2018.01.020

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz Jr., K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J., 2005a. The Amber biomolecular simulation programs. J. Comput. Chem. 26, 1668–1688. https://doi.org/10.1002/jcc.20290

Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J., 2005b. The Amber biomolecular simulation programs. J. Comput. Chem. 26, 1668–1688. https://doi.org/10.1002/jcc.20290

Cornwell, W., Nakagawa, S., 2017. Phylogenetic comparative methods. Curr. Biol. 27, R333–R336. https://doi.org/10.1016/j.cub.2017.03.049

Darden, T., York, D., Pedersen, L., 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J. Chem. Phys. 98, 10089–10092. https://doi.org/10.1063/1.464397

Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.-P., Simmonett, A.C., Harrigan, M.P., Stern, C.D., Wiewiora, R.P., Brooks, B.R., Pande, V.S., 2017. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLOS Comput. Biol. 13, e1005659. https://doi.org/10.1371/journal.pcbi.1005659

Ewald, P.P., 1921. Die Berechnung optischer und elektrostatischer Gitterpotentiale. Ann. Phys. 369, 253–287. https://doi.org/10.1002/andp.19213690304

Fischer, E., 1894. Einfluss der Configuration auf die Wirkung der Enzyme. Berichte Dtsch. Chem. Ges. 27, 2985–2993. https://doi.org/10.1002/cber.18940270364

Flechsig, H., Mikhailov, A.S., 2019. Simple mechanics of protein machines. J. R. Soc. Interface 16, 20190244. https://doi.org/10.1098/rsif.2019.0244

Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., Ferrin, T.E., 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. Publ. Protein Soc. 27, 14–25. https://doi.org/10.1002/pro.3235

Hagberg, A., Schult, D., Swart, P., n.d. Exploring Network Structure, Dynamics, and Function using NetworkX, in: Proceedings of the 7th Python in Science Conference. Presented at the SciPy 2008, G Varoquaux, T Vaught, J Millman, pp. 11–15.

Illergård, K., Ardell, D.H., Elofsson, A., 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins 77, 499–508. https://doi.org/10.1002/prot.22458

Jo, S., Kim, T., Iyer, V.G., Im, W., 2008. CHARMM-GUI: A web-based graphical user interface for CHARMM. J. Comput. Chem. 29, 1859–1865. https://doi.org/10.1002/jcc.20945

Koshland, D.E., 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc. Natl. Acad. Sci. U. S. A. 44, 98–104. https://doi.org/10.1073/pnas.44.2.98

Kufareva, I., Abagyan, R., 2012. Methods of protein structure comparison. Methods Mol. Biol. Clifton NJ 857, 231–257. https://doi.org/10.1007/978-1-61779-588-6_10

Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., Simmerling, C., 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput. 11, 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255

Parolini, G., 2015. The Emergence of Modern Statistics in Agricultural Science: Analysis of Variance, Experimental Design and the Reshaping of Research at Rothamsted Experimental Station, 1919–1933. J. Hist. Biol. 48, 301–335. https://doi.org/10.1007/s10739-014-9394-z

Petersen, H.G., 1995. Accuracy and efficiency of the particle mesh Ewald method. J. Chem. Phys. 103, 3668–3679. https://doi.org/10.1063/1.470043

Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., Ferrin, T.E., 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. Publ. Protein Soc. 30, 70–82. https://doi.org/10.1002/pro.3943

Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E., Ribeiro, J.V., Bernardi, R.C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M.C.R., Radak, B.K., Skeel, R.D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L.V., Schulten, K., Chipot, C., Tajkhorshid, E., 2020. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. 153, 044130. https://doi.org/10.1063/5.0014475

Pierce, L.C.T., Salomon-Ferrer, R., Augusto F. de Oliveira, C., McCammon, J.A., Walker, R.C., 2012. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. J. Chem. Theory Comput. 8, 2997–3002. https://doi.org/10.1021/ct300284c

Ponte, P., Melko, R.G., 2017. Kernel methods for interpretable machine learning of order parameters. Phys. Rev. B 96, 205146. https://doi.org/10.1103/PhysRevB.96.205146

Rajendran, M., Babbitt, G.A., 2022. Persistent cross-species SARS-CoV-2 variant infectivity predicted via comparative molecular dynamics simulation. R. Soc. Open Sci. 9, 220600. https://doi.org/10.1098/rsos.220600

Rajendran, M., Ferran, M.C., Babbitt, G.A., 2022. Identifying vaccine escape sites via statistical comparisons of short-term molecular dynamics. Biophys. Rep. 2, 100056. https://doi.org/10.1016/j.bpr.2022.100056

Ribeiro, J.V., Bernardi, R.C., Rudack, T., Stone, J.E., Phillips, J.C., Freddolino, P.L., Schulten, K., 2016. QwikMD — Integrative Molecular Dynamics Toolkit for Novices and Experts. Sci. Rep. 6, 26536. https://doi.org/10.1038/srep26536

Roe, D.R., Cheatham, T.E.I., 2013. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput. 9, 3084–3095. https://doi.org/10.1021/ct400341p

Rynkiewicz, P., Lynch, M.L., Cui, F., Hudson, A.O., Babbitt, G.A., 2021. Functional binding dynamics relevant to the evolution of zoonotic spillovers in endemic and emergent Betacoronavirus strains. J. Biomol. Struct. Dyn. 1–19. https://doi.org/10.1080/07391102.2021.1953604

Sali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815. https://doi.org/10.1006/jmbi.1993.1626

Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S., Walker, R.C., 2013. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. 9, 3878–3888. https://doi.org/10.1021/ct400314y

Sousounis, K., Haney, C.E., Cao, J., Sunchu, B., Tsonis, P.A., 2012. Conservation of the three-dimensional structure in non-homologous or unrelated proteins. Hum. Genomics 6, 10. https://doi.org/10.1186/1479-7364-6-10

Strong, M., 2004. Protein Nanomachines. PLoS Biol. 2, e73. https://doi.org/10.1371/journal.pbio.0020073

Suzuki, Y., 2010. Statistical methods for detecting natural selection from genomic data. Genes Genet. Syst. 85, 359–376. https://doi.org/10.1266/ggs.85.359

Tripathi, A., Bankaitis, V.A., 2017. Molecular Docking: From Lock and Key to Combination Lock. J. Mol. Med. Clin. Appl. 2, 10.16966/2575-0305.106.

Walker, R.C., Crowley, M.F., Case, D.A., 2008. The implementation of a fast and accurate QM/MM potential method in Amber. J. Comput. Chem. 29, 1019–1031. https://doi.org/10.1002/jcc.20857

Wang, H., Zhang, P., Schütte, C., 2012. On the Numerical Accuracy of Ewald, Smooth Particle Mesh Ewald, and Staggered Mesh Ewald Methods for Correlated Molecular Systems. J. Chem. Theory Comput. 8, 3243–3256. https://doi.org/10.1021/ct300343y

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A., 2004. Development and testing of a general amber force field. J. Comput. Chem. 25, 1157–1174. https://doi.org/10.1002/jcc.20035