**Energy modeling of the EGFR pathway**

**Abstract:**

**Introduction:**

The Epidermal Growth Factor Receptor (EGFR) signaling is a significant signal transduction pathway actively involved in the cellular development, proliferation, and differentiation regulations [1, 2]. It has been found to be major driver of many types of cancers, and neurodegenerative disease rendering it a subject of medical interest [3]. EGFR activation is induced by ligand bindings such as Epidermal Growth Factor (EGF), which in turn leads to EGFR dimerization and several tyrosine residues auto-phosphorylation. The phosphorylation in the cytoplasmic receptor domain initiates a biochemical communication between the receptor-ligand combination and cytoplasmic target proteins such as Src homology and collagen domain protein (Shc), growth factor receptor-binding protein 2 (Grb2), and phospholipase C-γ (PLCγ) [4, 5]. These early processes, including binding and phosphorylation of monomers using their kinase domains stimulate proliferation and differentiation, and thus result in short-term responses within the EGFR signaling cascade.

Kholodenko et al. studied this short-term behavior using a combination of a computational model and experimental data [5]. They developed a detailed kinetic model of early signaling events that considered a fairly large number of components and interaction parameters. On the experimental side, they showed it was possible to measure dynamics at fairly high temporal resolution using standard biochemical methods with fairly high precision, especially in comparison to most data that is reported in the literature. Despite a sophisticated mathematical model of the factors governing the kinetics of the EGFR signaling pathway, the model outputs do not adequately explain the transient behavior of several model components in experimental data. Furthermore, due to the model's manual incorporation of thermodynamic constraints, some detailed balance constraints have not been satisfied.

The question of how molecular diversity which arises from multi-domain protein-protein interactions influences the functional behavior of the EGFR signaling components has been addressed in the model developed by Blinov et al [6]. This model is mainly based on what Kholodenko et al. presented, however it takes the protein tyrosines into account separately. Nonetheless, the model of Blinov et al. still has some major shortcomings. Like the Kholodenko model, they do not use a systematic parameter estimation approach to find the best fit to the data. While they distinguish between EGFR phosphorylation and its binding sites, they do not consider the possibility of between-site cooperativity. Furthermore, both models specify rate constants directly, leaving open the possibility of violating thermodynamic constraints.

Rule-based modeling represents the interconnected molecules as structured objects which are governed by specific rules. Several protein-protein interaction consequences, including the formation of heterogeneous protein complexes, protein degradation, and post-translational protein modification, are present in nearly all biological process networks[7-9]. Typically, proteins' modular domains, which comprise binding and catalytic functions, mediate these processes. Despite the importance of these interactions on comprehending the dynamics of the biological system, adopting standard methods for considering these details generally leads to a complex network model. However, Rule-based modeling approaches enable modeling the site-specific details of protein-protein interactions while reducing combinatorial complexity [8].

Although a rule-based model provides a framework for demonstrating several reactions with a single reaction rule, it is constrained by the issues of detailed balance and regulatory complexity. The problem of detailed balance arises when reversible reaction rules are stated as though they are independently determined rather than mutually restricted by thermodynamic cycles [9]. Whereas cooperative interactions between the sites always lead to regulatory complexity issues, especially when multiple sites are used to govern various operations. Therefore, it takes numerous rules as well as, manually enforced detailed balance constraints to explain cooperative interactions and thermodynamic laws, which could result in a large model [9, 10].

Rule-based models can be specified in BioNetGen language (BNGL) for developing the modular and structure-based model of biochemical reaction networks. BioNetGen (BNG) is a rule-based software tool that uses a graph syntax and a set of reaction rules to define the molecular patterns and a large class of biochemical reactions, respectively [7, 8]. Moreover, it offers a concise framework for specifying a free energy accounting system based on the network generation methods for allosteric transitions. With the help of this BNG extension, known as eBNG, reaction rate laws may be calculated from reaction free energies and detailed balance constraints are automatically enforced. Furthermore, a system network with high order cooperative interactions could be handled efficiently by an energy-based model that is implemented in this formalism [9, 10].

In this study, we seek to rectify the weaknesses of the previous models by employing more systematic methodologies to explore the behavior of EGFR pathway. First, we attempt to map the relationships between cooperativities and the output of the system. In doing so, we are eager to gain a deeper understanding of how different model parameters influence signal transmission in EGFR pathway. Second, we aim to figure out what mechanism(s) cause transient in this system. To these ends, after demonstrating that the previously mentioned models are physically implausible due to violations of detailed balance, a rule-based version of the EGFR pathway model based on the Kholodenko’s paper is built in which detailed balance constraints are considered. Then, using an energy-based extension of BioNetGen (eBNG), we address the problem of regulatory complexity brought on by the cooperative interactions between different sites in the EGFR signaling pathway. Energy-based modeling has the advantage of enforcing thermodynamic restrictions automatically, as well as smaller number of parameters. Moreover, applying eBNG gives us the opportunity of easily expanding the model and bringing the model closer to reality by adding some interactions like simultaneous multiple binding of proteins to EGFR tyrosine residues. Finally, we apply parameter estimation methodologies to fit these versions of the EGFR signaling cascade model to the experimental data, and the results are used to infer a probable mechanism for transient production when it occurs.

**Methods:**

* **Building the rule-based model of EGFR signaling pathway**

The first presented model in this study is a rule-based version of the model in [5]. The aim of developing this model is not only providing the fundamental structure of energy modeling procedure, but also obtaining a set of optimal parameters utilizing parameter estimation techniques and imposing detailed balance constraints in order to have a fitted model of EGFR pathway against the experimental data which is believed to be plausible. The model, which is identical to that of Kholodenko et al. [5], has 50 parameters and 25 rules for creating the reaction network of the EGFR signaling cascade. However, 45 parameters are unknown as a result of the enforcement of detailed balance restrictions. More details on the model components, and biochemical reactions could be bound at [5]. The rule-based model was developed in BioNetGen which offers the opportunity of creating the multi-domain species resulting in a fewer number of model components [11].

**- The simple energy model**

In order to demonstrate some of the key features of the energy model, we first create a simple model containing a receptor, R, and two adaptors, A and B. The receptor could have two sites, one that binds A and one that binds B. Upon binding to R, A could become phosphorylated and B could bind to it. Using this simple model, We try to find parameters that would make it exhibit transient kinetics for A or B binding similar to what Kholodenko et al. [5] observed in their experiments. The model is developed both in rule-based and energy-based frameworks. These structures could be used to demonstrate how developing an energy model may aid in overcoming the challenges of detailed balancing constraints and regulatory complexity. It may also provide a simplified explanation of the potential mechanisms underlying EGFR transient behavior.

Here is a brief overview of the energy-based modeling principle and its implementation in BioNetGen. Each reaction should be primarily expressed in free energy space, where free energy variations () are associated to the equilibrium constants () as follows:

in which and are universal gas constant and temperature, respectively. could be scaled by . Therefore, reaction free energy is stated as the equation below:

and are the forward and backward reaction rates in a reversible chemical reaction, respectively [9]. In eBNG, a rate law function is expressed by Arrhenius theory of reaction rates for a reaction class containing all the reactions that share the similar rate law function. Then, and for any reaction in a class might be derived from the following equations based on linear transition state theory [12]:

In the above equations, denotes activation energy and is the rate distribution parameter which has a constant value between 0 and 1 and controls the contribution of in and [9].

The model in rule-based structure has 8 reaction rules including 14 reversible reactions and 1 non-reversible one, as well as 15 parameters. Due to the presence of one thermodynamic restriction along a cyclic set of reactions, there is a detailed balance constraint which reduces the number of independent reaction rates to 14. However, building the equivalent energy model needs only 5 reaction rules and 11 parameters. Besides having fewer number of model components, energy-based modeling framework eliminates the need to worry about enforcing the detail balance restrictions, because it implements this principle automatically.

…….The role of this simple model in describing the hypothesis about EGFR transient behavior……

* **Building the energy-based model in some steps**

The proposed energy model of EGFR signaling pathway in this study is developed in several steps. Each step involves the incorporation of one or more chemical species, as well as their related parameters and reactions. The model in the first step is comprised of three species: EGF, EGFR, and Shc. Therefore, the reaction rules define EGF binding to EGFR, EGFR dimerization, Shc binding to EGFR, transphosphorylation and dephosphorylation of tyrosines in Shc and the cytoplasmic portion of EGFR. Two more model elements, Grb2 and Son of Sevenless (SOS), are used in steps 2 and 3. Step 2 considers Grb2 binding to EGFR and SOS, while the adaptor role of phosphorylated Shc for enabling indirect Grb2-EGFR binding is modeled in step 3. In addition to the above mentioned interactions, PLCγ phosphorylation, activation, and binding to EGFR are the other early events which have been modeled in step 4. The kinetic parameters of the model at these four steps are adjusted so that the final model in step 4 recovers the model outputs in [5]. It's worth noting that the model in step 4 has substantially less components than the model in the Kholodenko et al.’s study [5], with only 19 reaction rules and 34 parameters.

Finally, the energy-based model is extended by discarding some restrictive assumptions in the Kholodenko et al.’s model [5]. This leads to a more realistic model by adding some interactions including simultaneous binding of multiple proteins to their target phosphotyrosine residues, and allowing the phosphorylated receptors to undimerize. Despite the addition of these interactions that were not taken into account in the original Kholodenko et al.’s model [5], the subsequent model in step 5 also has fewer parameters than the original model. The energy-based models in different steps can be found at: ….

- Model fitting

1- Parameter estimation and MCMC sampling using PEtab format and pyPESTO

The EGFR pathway models given in this study, including rule-based and energy-based models, are fitted to the experimental data in [5] using the python parameter estimation package pyPESTO [13]. It provides an algorithm for multi-start optimization and uncertainty analysis in computational biology problems which help to find the global optimum utilizing various optimization methods. For problem specification, we use pyPESTO linked to PEtab which is a standardized data format that defines measurement noise, parameter bounds, model outputs, and experimental conditions to prepare data for parameter estimation problems [14]. Using the applicable features of pyPESTO, we further evaluate the properties of the optimization problem using uncertainty analysis and Markov chain Monte Carlo sampling techniques.

- Sensitivity Analysis

**Discussion:**

Kholodenko et al. do not perform a rigorous fit of the model parameters to experimental data, leaving us without confirmation that their parameters provide the best fit to observation. Additionally, their model conflates all tyrosine phosphorylation sites on EGFR, without considering the possibility of multiple effectors binding at different sites on the same molecule.

1. Herbst, R.S., *Review of epidermal growth factor receptor biology.* International Journal of Radiation Oncology\* Biology\* Physics, 2004. **59**(2): p. S21-S26.

2. Oda, K., et al., *A comprehensive pathway map of epidermal growth factor receptor signaling.* Molecular systems biology, 2005. **1**(1): p. 2005.0010.

3. Murphrey, M.B., L. Quaim, and M. Varacallo, *Biochemistry, Epidermal Growth Factor Receptor.* 2018.

4. Batzer, A., et al., *Hierarchy of binding sites for Grb2 and Shc on the epidermal growth factor receptor.* Molecular and cellular biology, 1994. **14**(8): p. 5192-5201.

5. Kholodenko, B.N., et al., *Quantification of short term signaling by the epidermal growth factor receptor.* Journal of Biological Chemistry, 1999. **274**(42): p. 30169-30181.

6. Blinov, M.L., et al., *A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity.* Biosystems, 2006. **83**(2-3): p. 136-151.

7. Faeder, J.R., et al., *Rule‐based modeling of biochemical networks.* Complexity, 2005. **10**(4): p. 22-41.

8. Faeder, J.R., M.L. Blinov, and W.S. Hlavacek, *Rule-based modeling of biochemical systems with BioNetGen*, in *Systems biology*. 2009, Springer. p. 113-167.

9. Sekar, J.A., J.S. Hogg, and J.R. Faeder. *Energy-based modeling in BioNetGen*. in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2016. IEEE.

10. Ollivier, J.F., V. Shahrezaei, and P.S. Swain, *Scalable rule-based modelling of allosteric proteins and biochemical networks.* PLoS computational biology, 2010. **6**(11): p. e1000975.

11. Blinov, M.L., et al., *BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains.* Bioinformatics, 2004. **20**(17): p. 3289-3291.

12. Leffler, J.E., *Parameters for the description of transition states.* Science, 1953. **117**(3039): p. 340-341.

13. Stapor, P., et al., *PESTO: parameter estimation toolbox.* Bioinformatics, 2018. **34**(4): p. 705-707.

14. Schmiester, L., et al., *PEtab—Interoperable specification of parameter estimation problems in systems biology.* PLoS computational biology, 2021. **17**(1): p. e1008646.