Structure–activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era

Stephen V Frye

Extension of the traditional pharmacological approach of protein target classification to whole target systems has the potential to relate elements of protein sequence to the structure–activity relationship (SAR) of small molecules that can modulate protein action. Grouping potential drug discovery targets into families based on the relatedness of their SAR provides a means to translate the information from genome-sequencing efforts into knowledge that will aid in the discovery of drugs.

Address: Division of Chemistry, Glaxo Wellcome Inc., 3.4134, 5 Moore Drive, Research Triangle Park, NC 27709, USA.

E-mail: svf3511@glaxowellcome.com

Chemistry & Biology January 1999, **6**:R3–R7
http://biomednet.com/elecref/1074552100 6R0003

 ISSN 1074-5521

Structure–activity relationship homology (SARAH)

Structure–activity relationships (SARs) emerge from an analysis of how changes in small-molecule structure effect activity versus a particular molecular target. SAR for a set of small molecules can be either similar between protein targets, or distinct. The idea of affinity fingerprinting proteins with small molecules has been previously discussed as a method to select small-molecule sets for subsequent screening that are enriched in high-affinity compounds. The notion that affinity fingerprinting could also provide a quantitative measure of similarity between disparate proteins was also suggested [1]. Building upon this idea, targets that share SAR can be thought of as SAR homologous. SARAH is a convenient acronym for this relationship. This functional definition of relatedness between proteins has a long history in pharmacology. For example, the receptors of the sympathetic and parasympathetic nervous system were largely described and characterized prior to knowledge of their amino-acid sequence, and small-molecule agonist and antagonist activity was used extensively to classify these proteins [2]. With the advent of sequence information to group proteins into classes, this form of functional definition has become less widely used. The SARAH between proteins is uniquely valuable for assessing potential selectivity issues, however, when searching for potent and selective small molecules versus protein targets. Indeed, combinatorial chemistry and high-throughput screening provide the means to establish extensive empirically derived SARAH protein families. The resurrection of this classification technique is likely to be a very powerful tool in bioorganic chemistry and drug discovery.

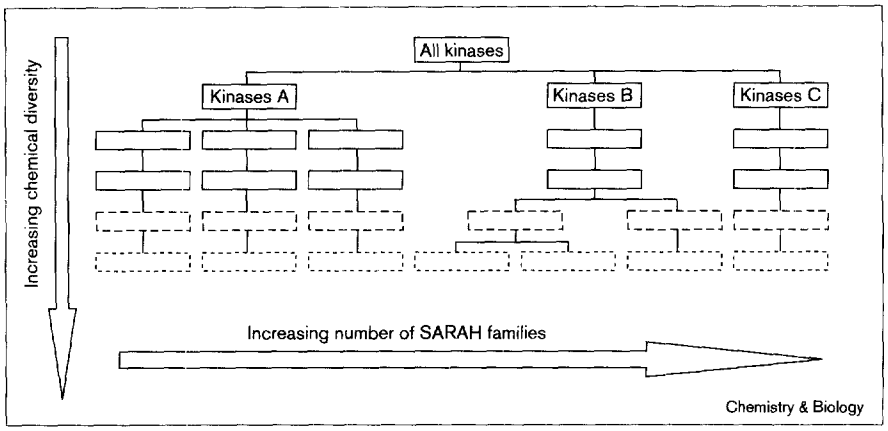
Systems-based research

The discovery of drugs was initially driven by a focus on physiology and empirical observations of the effect of small molecules in animal models of disease. Biochemistry and pharmacology then drove a phase of drug discovery based on understanding the molecular foundation for physiological processes. Medicinal chemistry programs could then focus around the knowledge of endogenous small-molecule mediators or enzymatic mechanism instead of *in vivo* empiricism. The last decade has seen an emergence of more systematic approaches to drug discovery through a focus on particular protein families that have proven amenable to intervention by small molecules: 7-transmembrane (7TM) receptors, nuclear hormone receptors, ion channels, proteases and protein kinases. The ability to systematically develop the molecular biology (cloning, expression and purification), biochemistry (screening) and medicinal chemistry within a protein target system allows for great efficiencies in all stages of preclinical drug discovery. This approach also allows scientific expertise to grow and flourish in an age where individual projects may only last for 2–3 years. Systems-based drug discovery starts with a target class of proven value in drug discovery and then seeks to use small molecules to make the connection to disease. I will refer to this form of research as 'systems-based research' [3]. The protein kinase family is such an emerging system for drug discovery.

The information challenge

Biomedical science is experiencing a flood of new information due to the ability of the technological advances in DNA manipulation and sequencing to rapidly provide the sequences of all expressed genes and entire genomes [4]. Within the context of drug discovery in the protein kinase arena, this information holds promise and challenges for the future. The primary sequence of all protein kinases in human and many pathogen genomes will soon be available, as well as some genetic disease association for many kinases or, more likely, the pathways in which they function. The existence of this information raises a number of questions: how will potential targets be prioritized for drug discovery efforts in a knowledge-based manner from among the estimated 3000 protein kinases? How can selectivity be assessed in a genomic context? (Currently

Figure 1



Sorting different kinases into SARAH families using small-molecule inhibitors. From an initial single family defined by ATP binding, the number of SARAH families increases as the chemical diversity of small-molecule kinase inhibitors increases.

many companies screen against 10–20 kinases for selectivity determinations; surely this is meaningless in such a large protein family.) Is there a conceptual framework that will allow us to rise to these challenges?

SARAH applied to protein kinases

Most protein kinases bind ATP. Ignoring differences in K_m and taking ATP binding as the sole discriminator, most kinases can be assigned to the same SARAH 'family'. The small-molecule structure (ATP) binding to their active site defines this family. (And indeed, many other proteins belong to this family based solely on ATP binding.) This is the minimal SARAH family for protein kinases, and certain conserved residues contribute to the binding of ATP and the catalytic mechanism of phosphate transfer. New family members can be picked out of the expressed sequence tag (EST) databases due to the sequence conservation that preserves this function. When we enter the world of small-molecule inhibitors, however, all protein kinases are no longer equivalent in terms of SARAH. With increasing diversity of inhibitors the previously irreducible SARAH family is split into subfamilies. This sorting of kinases into families by small molecules is entirely analogous to the traditional assignment of 7TM receptor families by their small-molecule agonists and antagonists and is similarly independent of any assumptions about the binding site of the small molecule. This concept is represented in Figure 1.

As it turns out, the vast majority of potent protein kinase inhibitors are, in fact, competitive with ATP, and selectivity arises from differences in pockets adjacent to the ATP-binding region that ATP itself does not occupy ([5] and references therein). Given that the sequence determinants of ATP binding are distinct from substrate- and protein-recognition determinants, we can anticipate that SARAH families based on ATP-competitive small molecules will not necessarily be coincident with biochemical families or overall sequence homology families (serine/threonine versus tyrosine, or receptor versus nonreceptor versus nuclear). In fact, these SARAH families will share binding function or 'shape' and perhaps sequence conservation around the ATP-binding site.

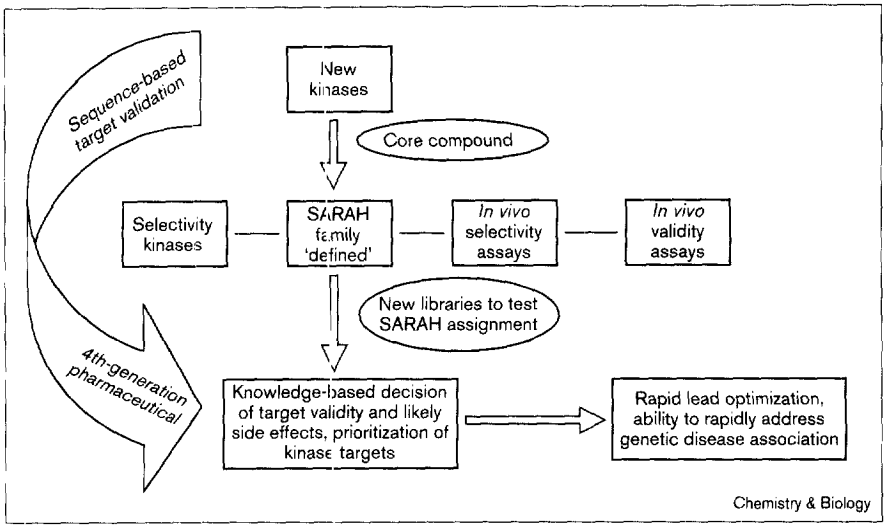
Potential applications of SARAH

By screening a common and diverse set of small molecule inhibitors against a set of protein kinases a SARAH assignment will be derived for each kinase [1]. Branch points in the family tree could be defined by overall correlation of pIC_{50} values between kinases. If a sufficient number of kinases is screened (undefined at this point), SARAH families will probably be populated by enough members to correlate amino-acid changes in the ATP-binding region (or ATP active-site shape based on X-ray crystallography and modeling) with the overall family assignments. Establishment of this correlation between sequence and small-molecule SAR becomes a 'Rosetta Stone' for translating protein-kinase sequence into the domain of drug discovery. Another important result of the SARAH family assignments is the fact that it provides a knowledge-based way to choose selectivity assays. Intelligent screening for potential cross-reactive kinases would examine members of the same SARAH family where, given the current state of chemical diversity, selectivity is anticipated to be a problem. In addition, representative kinases, especially ones implicated as important in physiology/toxicology, would be chosen for screening from other SARAH families in order to ensure that selectivity is maintained. Prioritization of kinase drug discovery targets could be based partially on selectivity/toxicity issues within this protein class as evidenced by the SARAH family to which any new target may belong. Potential targets that lie close in SARAH homology to known kinases whose inhibition is deleterious would be of lower priority than kinases whose SARAH assignment is less related to kinases to be avoided.

Although the SARAH family structure arises from small molecules sorting proteins into classes, the reciprocal

Figure 2

The drug discovery process flow using SARAH families. Once an initial core compound set has been created, it may be used to empirically classify any new kinase screened into a SARAH family. The SARAH assignment then provides the basis for establishing appropriate selectivity assays and active compounds from the core set may be useful for target validation work (assessment of connection of the target to disease of interest). The possibility of rapidly progressing many kinases to this level of understanding can aid in prioritization. When a correlation between sequence and SARAH can be made with confidence (curved arrow on far left of figure), sequence alone could provide the knowledge necessary for target prioritization.



relationship is also worth examining: kinases sort small molecules into clusters in pIC_{50} space that is dimensionally as large as the number of kinases screened ([1]; L.F. Kuyper, personal communication). Compounds are clustered due to their profile (correlated affinity) across the kinases screened. Within a cluster, compounds presumably share a similar shape/charge interaction across the current set of kinases screened. These compound clusters could fragment as biological diversity increases (i.e., more are kinases screened), in the same manner as previously irreducible SARAH families when chemical diversity increases. In analogy to the SARAH families' ability to correlate sequence with SAR, because of a shared active-site shape, molecules within a cluster share a shape that defines their activity across the set of kinases. At any point in the development of the biological and chemical diversity of this system it will be possible to select the compounds and the kinases that maximally differentiate each other. This core set (training set in [1]) then becomes the basis for reducing the number of compounds and kinases necessary for systematic exploration of new kinases and compounds. A representation of the idealized process flow within this conceptual framework appears in Figure 2.

What is chemical diversity anyway?

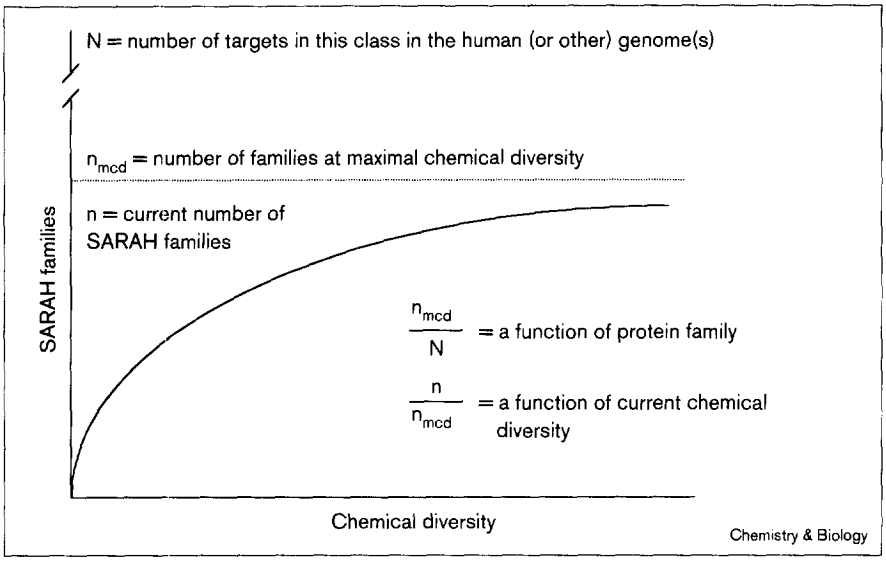
The interplay between chemical and biological diversity in this system is apparent in the previous discussion and it is of interest to ask how the number of SARAH families might relate to chemical diversity. Figure 3 depicts a purely speculative hypothesis of this relationship for protein kinases.

The hypothesis implied in Figure 3 is that as chemical diversity increases the number of SARAH families approaches a limit defined as n_{mcd} (number of families at maximum chemical diversity) and that this number will be less than N, the total number of protein kinases in the genome (or genomes). This seems likely in the kinase system because of the large number of members in this class and the fact that potent small-molecule inhibitors generally interact in the somewhat conserved ATP-binding site. If n_{mcd}/N is equal to 0.1 then a SARAH family would, on average, have ten members. Chemical diversity that explores other regions of the kinase could obviously change the assumptions behind this proposition. The model does, however, provide an operational definition of chemical diversity that is dependent upon the extrinsic properties of small molecules and the protein class targeted. In the realm of drug discovery, it seems that the most useful definition of 'chemical diversity' is that it is the property of a set of compounds which increases as the number of SARAH families increases. A correlation analysis between this empirical and extrinsic definition with the intrinsic properties of small molecules would seem to provide an appropriate basis for progress in the area of defining how chemical diversity varies with the intrinsic properties of molecules. (See [1] for an example using principle component analysis.) In the broader context of systems-based research, the ratio of n_{mcd}/N may be a fundamental characteristic of the protein family (or the particular site targeted), and methods to predict this characteristic would have great value for sorting the human genome into tractable classes of proteins for discovery of potent and selective drugs.

Initial results

Accumulation of a sufficient sized data set to begin to test the hypotheses resident in this conceptual framework is difficult even within a large pharmaceutical company. The concepts presented in this discussion are therefore

Figure 3



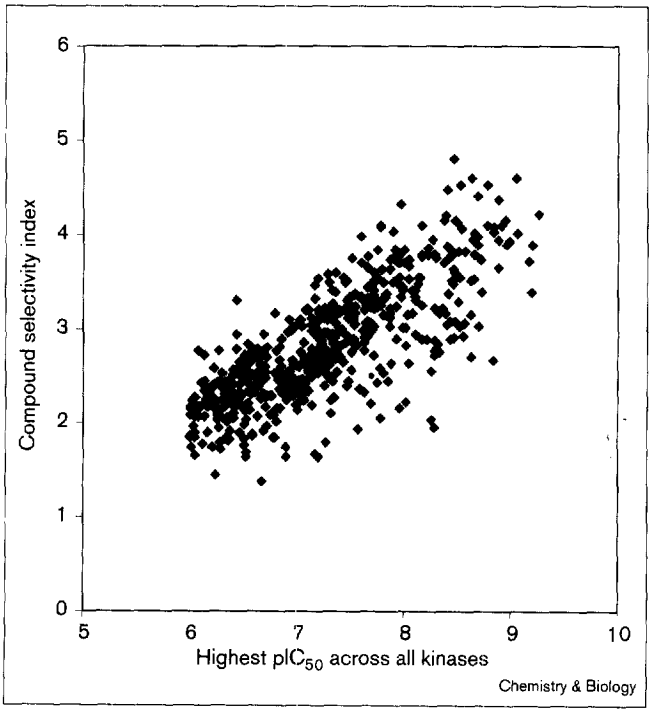
The interplay between chemical and biological diversity in the SARAH families. The number of SARAH families (n) increases as the diversity of small molecules screened increases. At some point kinases may be indistinguishable from each other with small molecule inhibitors. The number of SARAH families at the limit of maximum chemical diversity (mcd), n_{mcd}, compared to the total number of kinases in the genome (N) is expressed as the ratio n_{mcd}/N. This ratio is a fundamental property of a target class (or particular active-site targeted).

still mostly speculative at this stage. Some early results support the value of the general ideas, however. Figure 4 is a plot of compound selectivity index (defined as the highest pIC_{50} minus the average of all other pIC_{50} values) versus the highest pIC_{50} determined for that compound. The general correlation of potency and selectivity is consistent with the notion that more potent compounds are more finely tuned for the active site where they have a high affinity. This also means that the > 10 kinases against which this data was accumulated are somewhat distinct in their active-site shape. As the number of kinases screened versus this same set of compounds increases, the average selectivity of the compounds could increase, decrease or remain unchanged. Increasing selectivity at constant chemical diversity (i.e. the same compound set) is consistent with adding kinases of low SARAH to the current set of kinases. Decreasing selectivity indicates that the kinases added are of higher SARAH than the current set of kinases. There will be an analogous relationship as new compounds are added. More diverse compounds will increase the average selectivity, whereas less diverse will decrease it.

Translating information to knowledge

The general SARAH approach and related concepts discussed above are an attempt to build a framework in which chemistry and biology can most productively interface with the newer science of genomics [6]. The techniques of the yeast two-hybrid assay and differential gene expression can add significantly to this effort by bringing disease association and target validation information into this systems-based approach to protein kinases. The interrelationship of kinases based on protein partners and downstream genetic regulation is an important component of the 'Rosetta Stone' that can translate information into useful knowledge for the discovery of drugs. The combination of this systematic approach to SAR information with the power of X-ray crystallography and structure-based drug design also holds great promise for compound library design in the future.

Figure 4



The relationship between potency and selectivity for a set of small-molecule kinase inhibitors screened against > 10 protein kinases. Compound selectivity index is defined as the highest pIC_{50} minus the average of all other pIC_{50}s.

The questions posed at the beginning of this paper cut across many areas of biomedical research, especially in the pharmaceutical industry. The ability to create a meaningful interface between biology, chemistry and genomics, as presented here, could provide an entry into knowledge-based drug discovery in the context of the complete human genome.

Acknowledgements

The author recognizes Lee Kuyper, Mike Cory and Kevan Shokat for insightful discussions of the SARAH approach and the work of Kauvar *et al.* [1] as the foundation upon which SARAH rests.

References

1. Kauvar, L.M., *et al.,* & Rocke, D.M. (1995). Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **2**, 107-118.
2. Lefkowitz, R.J., Hoffman, B.B. & Taylor, P. (1990). Neurohumoral Transmission: the autonomic and somatic motor nervous systems. In *Goodman and Gilman's The Pharmacological Basis of Therapeutics*. (Gilman, A.G., Rall, T.W., Nies, A.S., Taylor, P., eds), Pergamon Press, New York.
3. Lehman, J., *et al., &* Williamson, R. (1996). Systematization of research. *Nature Supp.* **384**, 5.
4. Jasny, B.R. & Hines, P.J. (1998). A genome sampler. *Science* **282**, 651.
5. Toledo, L.M. & Lydon, N.B. (1997). Structures of staurosporine bound to CDK2 and cAPK – new tools for structure-based design of protein kinase inhibitors. *Structure* **5**, 1551-1556.
6. Schreiber, S.L. (1998). Chemical genetics resulting from a passion for synthetic organic chemistry. *Bioorg. Med. Chem.* **6**, 1127-1152.