

# Approaches to the detection of recessive effects using next generation sequencing data from outbred populations

David Curtis

Centre for Psychiatry, Barts and the London School of Medicine and Dentistry, London, UK

**Abstract:** Conventional methods to analyze genome-wide association studies and whole exome or whole genome sequencing studies would be prone to overlook variants which might exert a recessive effect on risk of disease, either as homozygotes or compound heterozygotes. It is plausible that such effects may be common even in outbred populations. An approach is described which is based on identifying a set of variants in a gene as being potentially of interest and then testing whether there is an excess of cases who are either homozygotes or complex heterozygotes for these variants. Methods based on departure from Hardy–Weinberg equilibrium are more powerful than those which compare cases to controls. However, linkage disequilibrium between variants can be difficult to deal with if phase is unknown. A simple approach for discarding variants apparently in strong linkage disequilibrium with others is proposed. The procedure is simple and quick to apply so can be used in the context of whole genome or exome sequencing studies and is implemented in the SCOREASSOC program.

**Keywords:** association, sequence, DNA

## Introduction

As discussed elsewhere in the context of schizophrenia,<sup>1</sup> it is entirely plausible that recessive effects might make a substantial contribution to susceptibility to disease in outbred populations but would not be detected by standard approaches to the analysis of genome-wide association studies (GWASs), single-nucleotide polymorphisms (SNPs) or next generation sequencing data. Although standard methods typically assume dominant or additive effects, an argument can be made that in terms of biological function one might well expect that random genetic variation would be more likely to act recessively. It is worth restating some of the arguments in favor of this view.

If a randomly occurring variant in a gene has an effect then it is likely to lead to a product which functions less well than the wild type. If a subject has two copies of the gene and the other is normal, then the usual outcome would be that there might be some overall lack of functioning due to a gene-dosage effect. Only rarely might a more specific mechanism produce a more dominantly acting effect; for example, by leading to an abnormal product which in some way interferes with the functioning of the normal product or is pathogenic through some other mechanism. Likewise, one might expect that it would be rare for a random variation to produce a gain of function. On the other hand, if both copies of the gene were abnormal, then one could readily arrive at a situation where there was marked or complete loss of function overall.

Another reason for suspecting that recessive effects may be common is through consideration of the effects of selection, which will be far weaker for recessive than

---

Correspondence: David Curtis  
Centre for Psychiatry, Barts and the London School of Medicine and Dentistry, London E1 1BB, UK  
Tel +44 20 7729 5207  
Fax +44 20 7729 5201  
Email david.curtis@qmul.ac.uk

dominant variants. To take the most extreme example, a dominantly acting variant which is lethal in youth will undergo complete selection and by definition can only arise through de-novo mutation. By contrast, variants which are lethal in homozygotes can persist in the population at relatively high frequency, and there are countless Mendelian recessive disorders which illustrate this. For recessively acting variants which increase susceptibility to disease without necessarily completely compromising reproductive fitness, selection pressures will exert even less effect on allele frequency. Recessive X-linked disorders represent an intermediate situation where selection occurs primarily through males. (Though there may be some secondary effects on female reproductive fitness; for example, if caring for an affected son impairs the ability to produce further children or if family stigma makes it more difficult for carrier sisters to marry and have children.)

If we acknowledge that there are *a priori* reasons to suspect that recessive effects might be common, we should go on to point out that many would have been overlooked by standard analytic methods. While there are examples of recessive mutations which have been detected in sequencing studies<sup>2,3</sup> it can be argued that standard approaches applied to outbred populations might overlook variants making an important contribution to risk, especially if they involved variants which were not very obviously damaging and/or formed compound heterozygotes. Before proceeding to discuss compound heterozygotes, we will first consider the simple situation where subjects homozygous for a single variant have increased risk of disease. With a Mendelian effect and no heterogeneity, all affected subjects would be homozygotes, but for more complex diseases, one would expect that only a proportion of cases would be homozygous for a given variant, since others would be due to the effects of other variants in the same gene, in different genes, or through other risk factors. There would thus be an increase in the number of cases homozygous for the risk variant, and there would be some increase in the overall frequency of this allele among cases, but this might be fairly modest. If one tested for it explicitly and if the variant were genotyped directly, then one could observe the excess of homozygotes, but in the context of a GWAS it is possible that tagging SNPs would not pick this up through being in incomplete linkage disequilibrium with the causal variant and/or through having a higher allele frequency. Thus, there might not be a marked tendency for the tagging SNPs themselves to be homozygous more frequently in cases than controls. In the context of a sequencing study, the excess of homozygotes

would be apparent if it was looked for specifically but might well be missed if a simple burden test was performed. It is also important to point out that recessively acting variants tend to have higher allele frequencies than dominant ones and hence they might well be overlooked in a sequencing study that tends to focus attention only on very rare variants.

The situation becomes more complex if we consider the possibility of different variants forming compound heterozygotes to produce a joint recessive effect on pathogenesis. This might well be a common situation, especially because such variants would be subjected to little in the way of selection pressure. If there were several different variants which could act together in this way then they could well be overlooked completely. In the context of a GWAS, each might be in linkage disequilibrium (LD) with different combinations of tagging SNPs, rendering it virtually impossible to detect the influence of any of them. In a sequencing study there would be only a modest increase in the frequency of each variant, and indeed of all variants combined. There would be expected to be some increase in homozygotes among cases, but in fact compound heterozygotes would occur more often than homozygotes, and any method which considered each variant individually, without evaluating the extent to which different variants co-occurred in the same subject, would have marked loss of power. The magnitude of these effects was quantified using a range of simple models.<sup>1</sup> To take one of these examples, we can consider a disease which has a population prevalence of 1%, with 5% of these cases being due to combinations of rare variants in a single gene acting recessively. If we assume that there are many rare variants in the gene but that only 25% of these have a pathogenic effect such as to increase the risk of disease to 0.2 if they occur as homozygotes or compound heterozygotes then a standard burden test will have power of only 3% to detect association at  $p = 10^{-6}$  with a sample size of 2000.<sup>1</sup> However, if a test is performed to specifically detect an excess of cases possessing at least two variants (either homozygotes or compound heterozygotes), then the power increases to 26%.

This paper explores some of the issues around developing methods which aim specifically at detecting such recessive effects.

## Methodology

### Single variants

Methods to detect an excess of homozygotes for a single variant are relatively straightforward. However, it is worth reiterating a point made elsewhere, which is that the test for an excess of homozygotes against the expectation under

Hardy–Weinberg Equilibrium (HWE) is far more powerful than the test for an excess of homozygotes among cases against controls.<sup>1</sup> To take the example described above, the power to detect an excess of homozygotes against cases and controls is 26%, but the power to detect excess homozygotes among cases compared with the number expected under HWE is 72%.<sup>1</sup> Similar substantial gains of power were found across a range of different recessive transmission models. An additional practical advantage of testing for departure from HWE is that one does not in fact need the control genotypes. One can simply use allele frequencies from the cases and controls to get a combined estimate of allele frequency and hence to determine the expected number of homozygous cases. If control allele frequencies are not available, then one can simply test for departure from HWE among cases. The appropriate test to use is a simple binomial test, which can be approximated by a Chi-squared test with one degree of freedom if expected counts are not too small.

While tests for departure from HWE are more powerful, they may be more susceptible to technical artefacts if homozygote and heterozygote calls cannot be reliably distinguished. This is not necessarily a major problem for comparisons of homozygote frequencies between cases and controls if there is no expectation that incorrect calls will be more frequent among cases.

In order for tests for departure from HWE to be valid, the assumptions underlying the expectation of HWE also need to be valid – essentially that the genotypes are drawn from a homogeneous sample in which random mating occurs. If these assumptions are violated to a substantial degree then observed departures from HWE might not necessarily provide evidence for recessive effects.

## Selection of multiple variants

If we wish to consider multiple variants jointly then we will be testing the hypothesis that some subset of these variants impair the functioning of the gene and hence can act recessively and increase risk of disease when one occurs as a homozygote or when two occur in *trans* to form a compound heterozygote.

There are a few topics which need to be addressed regarding which variants should be included in such an analysis. Firstly, one needs to define a region of interest. As a starting point, one might say that the unit of analysis should be a gene, but precisely what is meant by that would require some clarification.<sup>4</sup> For example, if we are considering only recessive effects, it would make sense to say that we would only be interested in pairs of variants which both damage

the same transcript and pairs which did not fulfill this condition would be irrelevant. One needs to consider the extent to which one would or would not include promoter regions and untranslated regions. One would also need to consider what types of variants to include. If the general approach will be to test whether there is an excess of cases possessing two or more variants then one needs to define which variants to include in a way which will mean that this condition would be met for only a small proportion of subjects under the null hypothesis, perhaps 10% or fewer. Thus, one could not include in the analysis very large numbers of variants or variants with a high minor allele frequency (MAF) because then all or nearly all subjects would possess at least two of them. In terms of function, one would probably not want to include synonymous or intronic variants. Although one would not want to include variants with high MAF, it is important to realize that variants having a recessive effect may have a considerably higher MAF than those with a dominant effect, and so the typical thresholds applied when searching for dominantly acting variants would not be appropriate. Certainly, one should not confine one's attention to only "novel" or extremely rare variants. The exact threshold chosen might depend on the specific circumstances, but a critical value of  $MAF < 0.1$  might be appropriate for situations where one is dealing with a relatively common disease (prevalence around 1%). If the homozygote had a penetrance of 0.1, then cases due to homozygotes of this variant would then have prevalence of 0.1%. Overall, we would want to select a set of variants based on position, function, and MAF such that 10% or fewer of subjects would be expected by chance to carry two or more variants.

## Detecting an excess of homozygotes across multiple variants

It is straightforward to carry out a simple  $2 \times 2$  Chi-squared analysis to see if cases are more frequently homozygous for at least one of the variants than controls.

However an alternative approach is to test for departure from HWE in cases. As noted above, this is more powerful than comparing the cases with controls. To derive a test statistic for this, we will begin by making the simplifying but usually false assumption that no LD exists between variants. Then, if we write  $p_i$  for the MAF of variant  $i$ , the probability for a case to be homozygous at this variant is  $p_i^2$ , and the probability to be homozygous for one or more variants is  $e_{\text{HOM}} = 1 - \prod (1 - p_i^2)$ . The MAF can be estimated jointly from cases and controls if available or if necessary only from cases. The observed number of cases who are homozygous

can then be compared with the expected number,  $e_{\text{HOM}} \cdot N_{\text{CASE}}$ , using a binomial test or, if the expected number is more than 5, a Chi-squared test. If the assumption of independence is violated, and LD is in fact present between variants, then the effect is to make this test more conservative. This is because if two variants tend to occur together, there is less opportunity for them to independently form homozygotes, and so the overall null hypothesis probability for a subject to be homozygous for at least one variant is reduced.

## Detecting an excess of homozygotes and compound heterozygotes

It should be pointed out that there may be considerable benefits in attempting to detect an excess of compound heterozygotes rather than only homozygotes. Compound heterozygotes are expected to occur more frequently than homozygotes, especially when variants are individually rare. If we consider two rare variants each with MAF  $p$ , then the population frequency for either homozygote is  $p^2$ , producing an overall frequency of  $2p^2$  for homozygotes, but the frequency for compound heterozygotes is  $4p^2(1 - p)^2$ . (We are ignoring all situations in which a subject has more than two variant alleles.) Thus, we could expect that there will be twice as many compound heterozygotes as there are homozygotes. This effect becomes stronger with larger numbers of variants.

Once again, it is straightforward to compare the counts between cases and controls of subjects carrying two or more of a set of variants. Although such a test remains valid if LD is present between variants (assuming it occurs to the same extent in both groups), LD can render the test problematic if it means that two variants frequently occur together and hence large numbers of subjects possess the same pair of variants. If a large proportion of controls carry at least two variants, then it becomes difficult or impossible to detect any excess among cases. Approaches to dealing with LD between variants will be discussed subsequently.

If for now we again make the assumption that there is no LD between variants, then we can derive the null hypothesis probability under HWE that a case will carry at least two variants. This is best approached iteratively by considering each variant in turn and obtaining the cumulative probabilities of a subject to have 0, 1, or more variants by the time we have included the  $i$ th variant. We will denote these cumulative probabilities as  $pAA_i$ ,  $pAB_i$  and  $pBB_i$ . For the first variant we write:

$$pAA_1 = (1 - p_1)^2$$

$$pAB_1 = 2(1 - p_1)p_1$$

$$pBB_1 = p_1^2$$

Then we incorporate the effect of the second variant and write:

$$pAA_2 = (1 - p_2)^2 \cdot pAA_1$$

$$pAB_2 = 2(1 - p_2)p_2 \cdot pAA_1 + (1 - p_2)^2 \cdot pAB_1$$

$$pBB_2 = p_2^2 \cdot pAA_1 + 2(1 - p_2)p_2 \cdot pAB_1 + (1 - p_2)^2 \cdot pBB_1$$

That is, for each new variant considered we take the probabilities that it increases the overall variant count from 0, 1 or more by 0, 1 or 2, so that for the  $i$ th variant we write:

$$pAA_i = (1 - p_i)^2 \cdot pAA_{i-1}$$

$$pAB_i = 2(1 - p_i)p_i \cdot pAA_{i-1} + (1 - p_i)^2 \cdot pAB_{i-1}$$

$$pBB_i = p_i^2 \cdot pAA_{i-1} + 2(1 - p_i)p_i \cdot pAB_{i-1} + (1 - p_i)^2 \cdot pBB_{i-1}$$

As before, we can now use a binomial or Chi-squared test to compare the observed number of cases carrying two or more variants, with the expected number  $pBB_{\text{ALL}} \cdot N_{\text{CASE}}$ .

Of course, if an overall excess of cases with two or more variants is detected, then it may not be obvious which particular variants are pathogenic. Once one had established the general effect, one might seek to identify which variants were driving it, but potentially there might be a large number of variants and pairs of variants to select from.

## Dealing with linkage disequilibrium between variants

Some preliminary investigation on real, though unpublished, datasets has revealed that it is by no means uncommon to find that variants within a gene can be in strong LD with each other. This can be observed even for variants which are extremely rare. The mechanisms which produce LD in such circumstances are not clear, because it might be assumed that it would be quite unlikely for one extremely rare variant to occur on the background of another. If LD between variants is not handled in some way, then the test for an excess of two or more variants based on HWE becomes completely invalid. If a pair of variants always occurs together because of LD, then if each subject possessing them is wrongly counted as a compound heterozygote, it would obviously produce a total number of subjects carrying two or more variants far higher than the expectation under independence outlined above.

The ideal solution to this problem would be if the sequencing method reported whether two variants were in *cis* or *trans*



and then only pairs of variants in *trans* would be counted as compound heterozygotes. If such information is not available, there are some statistical approaches which can be applied so that phase-unknown genotypes can be analyzed, albeit with some caveats.

The conventional approach to deal with this situation would be to use measures of LD or formal haplotyping algorithms to identify pairs of variants in LD with each other and then to remove one member until left with a set of independent variants. However, there are problems in applying these conventional approaches to the kind of data produced by sequencing studies because one has to deal with large numbers of variants, some of which are extremely rare. With typical sample sizes, there may be the situation where a subject carries two variants, neither of which is seen in any other subject. These would have a  $D'$  of 1, and any maximum likelihood method would assign them to the same haplotype and hence one of the pair would be discarded. If this approach was followed, one could end up wrongly ignoring many pairs of variants which did in fact represent compound heterozygotes. One might take the approach of attempting to estimate LD relationships only from control subjects, but this could be problematic if there were many rare variants, and if a substantial proportion only occurred in cases.

Methods to phase rare variants are being developed,<sup>5</sup> and as more sequence data becomes available then LD relationships between variants will become better understood, and it may be possible to assign even quite rare variants to known haplotypes. For now, the question remains as to the best way to utilize data arising from sample sizes running into the hundreds or low thousands, and it does seem that the datasets themselves can provide some useful information. If there is an excess of cases carrying at least two variants, then there are a few characteristics of the genotypes which may support the hypothesis that these represent compound heterozygotes rather than alleles in LD with each other. These characteristics being as follows:

1. If there is also an excess of homozygotes, then this is consistent with a recessive effect.
2. If there is no excess of controls carrying at least two variants.
3. If the variants observed to occur together also occur individually. If they occur in several subjects but always in the same pair, then this would suggest LD. If they sometimes occur together and sometimes not, then it can be difficult to determine whether they may be in LD to some extent.

4. If pairs of different variants are not observed together as double homozygotes. If double homozygotes do occur, then the haplotype is defined and LD can be assumed to be present.
5. If multiple different kinds of pairing are observed between different variants. If compound heterozygotes increase risk of disease, then we expect different pairs of variants to occur randomly among cases alongside homozygotes. However if we repeatedly observe the same pair, then we should suspect that LD is the explanation.

Although we can state these principles as being helpful to distinguish compound heterozygotes from variants in LD with each other, it is a challenge to incorporate them into a formal statistical test. At present there is no clear solution which is wholly satisfactory, although the situation is likely to improve in the future as more data are accumulated and more sophisticated statistical processes are developed.

For now, it is possible to propose a simple ad hoc approach, which is to discard a variant if it occurs more than once and if when it does occur there is one or more other variants which it “usually” occurs with. If quite a high value is chosen for “usually,” say 0.9 to only exclude pairs which almost always occur together, then there will be only a low risk of wrongly discarding any genes which might be of interest. For those genes which do then demonstrate a statistically significant excess of cases carrying two or more variants, one can then examine the genotypes of these subjects more closely. At this point a number of factors may be brought in to attempt to decide whether the results are of interest. These factors will include not only the variant genotypes themselves but also the nature of the gene, the predicted effect, frequency and position of the variants, whether there is any evidence for dominant effects, and so forth.

To summarize, the proposed procedure to detect whether recessively acting variants may contribute to the effect of a gene on disease susceptibility is as follows:

1. Define a region of interest; for example, all the exons of a gene or of one transcript, with or without untranslated regions, promoter regions, etc.
2. Define which variants to include; for example, nonsense, missense, frame-shift coding and splice site variants having  $MAF < 0.1$ . The aim should be that 10% or fewer controls will have two or more such variants.
3. Eliminate any variants which show strong LD with another variant; for example, where the second variant occurs in 90% of subjects carrying the first variant.

4. Test whether the proportion of cases homozygous for at least one of these variants is larger than the proportion of controls.
5. Test whether the proportion of cases homozygous for at least one variant is higher than would be expected assuming HWE. This finding will carry more weight if departure from HWE is not observed in controls.
6. Test whether the proportion of cases carrying two or more variants, ie, homozygotes or possible compound heterozygotes, is higher than the proportion of controls.
7. Test whether the proportion of cases carrying two or more variants is higher than would be expected under independence and HWE. Again, this finding will be more convincing if no excess is observed in controls.

This procedure has been implemented in the SCORE-ASSOC program, which also implements a weighted burden test to detect dominantly acting effects.<sup>6</sup> The tests are quick to perform and can readily be applied to exome-wide or genome-wide next generation sequencing results on case control samples involving hundreds or thousands of subjects.

It is not expected that the application of this process in isolation will produce conclusive evidence that a particular gene is or is not involved in the etiology of a disease. Rather, it can be applied to whole genome data to focus attention on a small subset of genes which seem to provide some support for recessive effects. Then, these genes can be examined in more detail. This will involve consideration of the plausibility of the gene, the observed pattern of variant genotypes, the predicted effects of the variants, whether there is also evidence that other variants in the gene produce dominant effects, and so forth.

## Discussion

An approach for detecting recessively acting variants is outlined. It is emphasized that unless special steps are taken to test for such variants, they may well be overlooked by conventional approaches to analysis. The approach is based on considering multiple variants within a gene jointly. Statistical tests which detect departure from HWE are more powerful than those based on comparisons of cases and controls, but when compound heterozygotes are considered, they are very sensitive to LD between markers. Hence, it would be desirable to have a means to identify variants which are in *trans* rather than *cis*, but in the absence of this information, attempts can be made to detect LD from the distribution of variant genotypes and then to go on to discard variants which appear to be in

strong LD with others. Application of such approaches may serve to focus attention on genes which might otherwise have been missed.

Although a test for departure from HWE in cases may be more powerful than testing for differences in genotype frequencies between cases and controls, this approach will only be valid if the variant alleles considered occur independently. Departures from this assumption might be caused by LD or population stratification. It is natural to look to the control genotypes to assist in judging whether such confounding factors are relevant, but if one does this then one is essentially returning to a test which compares case and control genotypes. There is thus a tension between gaining power through making assumptions about the independence of variant alleles in cases and relinquishing some power through making fewer assumptions and simply comparing cases with controls.

It is perhaps worth pointing out that unfortunately following up variants which potentially have a recessive effect may be more problematic than following up potential dominant variants. When recessive effects occur there may be a large number of variants which can combine to have a recessive effect on risk. If none tended to predominate, then when carrying out follow-up studies in additional samples one might need to carry out large numbers of genotyping assays to ensure that important effects were not missed. Alternatively, it might be necessary to sequence relevant regions of the gene in order to characterize rare and novel variants. Likewise, functional studies might be more challenging as they would need to assess the effect of homozygous variants or two variants being present in *trans*.

It should also be pointed out that one unsatisfactory property of the proposed procedure is that it adopts an all or nothing approach to deciding which variants should be included in the analysis. It has been argued that schemes which weight variants according to rarity and functional effects are preferable to those which arbitrarily include some variants but not others, and methods have been developed which incorporate such weighting schemes for dominantly acting variants.<sup>6-9</sup> However, it is not obvious how weighting methods should best be applied in the context of detecting recessive effects. An alternative to treating variants differently according to such factors as their rarity and predicted function would be to adopt a more post hoc, model-fitting approach which would attempt to make inferences about the nature and magnitude of the effect of each variant, though again it might be challenging to implement this in practice. Likewise, the proposal simply to discard some variants based

on apparent LD might not be optimal, and again it might be desirable to develop some scheme which could provide a more quantitative treatment.

It is hoped that if attention can be focused on this issue, then improved statistical and sequencing methods can be developed to more robustly characterize recessively acting variants, since they may make a substantial contribution to disease liability which would not be detected by methods which inherently focus on dominant effects.

## Acknowledgments

This study made use of data generated by the UK10K Consortium, derived from the following sample sets: Edinburgh schizophrenia samples, National Institute for Health and Welfare (THL) Finnish schizophrenia families from the Genetic Etiology of Severe Mental Disorders in Finland study, the Maudsley family study, the GAP study, the Maudsley twin study, CardiffScz, Scottish schizophrenia cases, University College London schizophrenia family samples, and the Severe Childhood Onset Obesity Project (SCOOP) UK. A full list of the investigators who contributed to the generation of the data is available from <http://www.UK10K.org>. Funding for UK10K was provided by the Wellcome Trust under award WT091310.

## Disclosure

The author declares he has no conflict of interest.

## Availability

The SCOREASSOC program, which implements the procedures described to detect recessively acting variants, is available as part of the GCPROG package at <http://www.smd.qmul.ac.uk/statgen/dcurtis/software.html>. C and C++ source code and a Windows executable are provided.

## References

1. Curtis D. Consideration of plausible genetic architectures for schizophrenia and implications for analytic approaches in the era of next generation sequencing. *Psychiatr Genet*. 2012;23(1):1–10.
2. Becker J, Semler O, Gilissen C, et al. Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. *Am J Hum Genet*. 2011;88(3):362–371.
3. Najmabadi H, Hu H, Garshasbi M, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature*. 2011; 478(7367):57–63.
4. Curtis D, Vine AE, Knight J. A simple method for assessing the strength of evidence for association at the level of the whole gene. *Adv Appl Bioinform Chem*. 2008;1:115–120.
5. Menelaou A, Marchini J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*. 2012;29(1):84–91.
6. Curtis D. A rapid method for combined analysis of common and rare variants at the level of a region, gene, or pathway. *Adv Appl Bioinform Chem*. 2012;5:1–9.
7. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009;5(2):e1000384.
8. Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet*. 2011;88(3):283–293.
9. Yi N, Liu N, Zhi D, Li J. Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. *PLoS Genet*. 2011;7(12):e1002382.

### Advances and Applications in Bioinformatics and Chemistry

Dovepress

### Publish your work in this journal

Advances and Applications in Bioinformatics and Chemistry is an international, peer-reviewed open-access journal that publishes articles in the following fields: Computational biomodeling; Bioinformatics; Computational genomics; Molecular modeling; Protein structure modeling and structural genomics; Systems Biology; Computational Biochemistry;

Computational Biophysics; Chemoinformatics and Drug Design; In silico ADME/Tox prediction. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <http://www.dovepress.com/advances-and-applications-in-bioinformatics-and-chemistry-journal>