Genomic Mappings and Spectral Analysis in The Frequency Domain

Sergey Edward Lyshevski
Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology
Rochester, NY 14623, USA
E-mail: Sergey.Lyshevski@mail.rit.edu Web: http://people.rit.edu/seleee

ABSTRACT

Departing from statistical methods, we examine large-scale genomic and proteomic data by applying spectral estimates and measures such as energy, power and cross-spectral densities. The frequency analysis and spectral methods have significant advantages, guarantee robustness, enable consistent quantitative analysis and provide qualitative features. The symbolic and numeric approaches provide the overall coherency. The frequency-domain analysis necessitates one to use numeric mappings of finite sequences. Though additional studies and consistent evaluations are needed to assess the proposed methodology, we demonstrate promising consistency, data cohesiveness as well as the genomic and proteomic correlations. Regression analysis and classifications can be achieved under large uncertainties (gaps, errors, missing sites, inconsistency, etc.). The analysis of sequences and information complexity requires a great number of assumptions, hypotheses and postulates. We minimize the number of assumptions applied. The results are illustrated for HIV, cancer and other sequences.

Keywords: Fourier transform, genome, genomics, energy spectral density, power spectral density, proteomics

1. INTRODUCTION

The large-scale genomics and proteomics are the forefront of medicine, life science and engineering. Current developments in genomics and proteomics promise one to identify and characterize genes, RNA, and proteins thereby enabling system biology, biotechnology and medicine [1, 2]. It is important to understand how genotype leads to phenotype, how an organism responds to the environment, and other open questions. Different statistical methods were used to analyze and evaluate large-scale data by performing data analysis and data mining. These attempts were partially successful due to overall complexity, sequences gaps, nonconding *low complexity* regions, inaccuracy, etc. The use of statistical methods in analysis of large-scale data, produced by high-throughput experiments, has limitations and drawbacks [1, 2].

Genome sequences for different organisms are available. In particular: (1) GenBank (USA), DDBJ (Japan) and EMBL databases provide nucleic acid sequences; (2) PIR and SWISS-PROT databases report protein sequences; (3) Protein Data Bank providesrs protein structures. Enabling databases were developed. For example, the SCOP, CATH and FSSP databases classify proteins based on structural similarity. The protein families were identified based on sequence homology applying Pfam and ProtoMap classifiers. The PartList and GeneCensus databases and classifiers examine the occurrence of protein families in various genomes.

Statistical methods test *a priori* hypotheses against data with a great number of assumptions and postulates under which the genome-genome comparison can be performed. The "learning" methods (clustering, Bayesian networks, decision trees, neural networks and other) were used to study trends and patterns in the large-scale data within moderate progress [1, 2]. We propose a frequency-domain approach which leads to the spectral analyses with consistent estimates and measures. This concept promises to ensure robust, systematic and consistent analysis [3, 4]. The proposed approach complies with conventional data formats and complements other methods ensuring assessment of complex large-scale data under uncertainties. The qualitative and quantitative analyses are performed for various sequences, including HIV and cancer genomes.

2. SPECTRAL ANALYSIS AND ITS APPLICATION

To perform analysis of various sequences we formulate the following postulate.

Postulate. *The quantifying data* (*relative information content*) *is coded, and, the descriptive features are defined as a finite sequence of nucleotides or amino acids in the genomic and proteomic sequences. These finite ensemble sequences are distinguishable and provide unique characteristics on identifiable quantities. Hence, the frequency-domain concept with the resulting spectral estimates and measures is applicable and consistent*. ■

Let $\mathbf{A}=\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ is the *symbolic quaternary alphabet*. This *alphabet* can be mapped (represented) as

$\mathbf{M}=\{0\ 1\ 2\ 3\}$, $\mathbf{M}=\{j\ \ -j\ \ 1\ \ -1\}$, $\mathbf{M}=\{1+j\ \ -1+j\ \ 1-j\ \ -1-j\}$.

Other mappings can be used. The arbitrary pairs of quaternary N-sequences (words of length N) are

$x=(x_1,x_2,\ldots,x_{N-1},x_N), x_i\in\mathbf{A}$ and $y=(y_1,y_2,\ldots,y_{N-1},y_N), y_i\in\mathbf{A}$.

For a pair (x,y) of quaternary words, the statistical measures and similarity $S(x,y)=\sum_{i=1}^{N} s(x_i,y_i)$ and entropy can be found. For N objects (symbols) X_i which have probability distribution functions $p(X_i)$, the entropy is $H(X)=-\sum_{i=1}^{N} p(X_i)\log_2 p(X_i)$, $i=1,2,\ldots,N-1,N$. We depart from these approaches which have a limited practicality.

Consider a finite sequence of nucleotides A, T, C and G. We assign the symbol or values a, t, c and g to the

characters A, T, C and G. These *a*, *t*, *c* and *g* can be mapped by real and complex mappings. The numerical sequence, resulting from a character string of length *N*, is

$x[n]=au_A[n]+tu_T[n]+cu_C[n]+gu_G[n]$, $n=0,1,2,\ldots,N-1$,

where $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are the binary indicators which take the value of either 1 or 0 at location *n* depending on whether the corresponding character exists or not at location *n*; *N* is the length of the sequence.

The amino acid sequences are expressed as

$x[n]=A_{la}u_{Ala}[n]+A_{rg}u_{Arg}[n]+\ldots+T_{yr}u_{Tyr}[n]+V_{al}u_{Val}[n]$.

Using the amino acids, the *symbolic alphabet* is **A**={Ala,Arg,…,Tyr,Val} with the corresponding *alphabet mapping*. The amino acid sequence is

$x[n]=au_a[n]+ru_r[n]+\ldots+tu_t[n]+vu_v[n]$.

We obtain the symbolic strings which map nucleotides and amino acids finite sequences. The discrete Fourier transform of a sequence $x[n]$ of length *N* is

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi}{N}kn},\ k=0,1,2,\ldots,N-1.$$

This Fourier transform provides a measure of the frequency content at frequency *k* which corresponds to a period of *N/k* samples. The sequences $U_A[k]$, $U_T[k]$, $U_C[k]$ and $U_G[k]$ are the discrete Fourier transforms of the binary indicators $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$. The finite genomic and proteomic sequences are distinguishable and may provide unique characteristics identifiable and observable in the frequency domain. These characteristics and data may not be observable by statistical,"learning" or other methods.

The energy spectral density (ESD) and power spectral density (PSD) associate with a stationary stochastic and deterministic functions and sequences. The aforementioned estimates are real-valued functions which represent the frequency content and identify periodicity.

The energy spectral density describes how the energy (or variance) of $x(t)$ or $x[n]$ vary as a function of frequency. If $x(t)$ is a finite-energy square-integrable functions, the spectral density $\Phi(\omega)$ of $x(t)$ is the square of the magnitude of the continues Fourier transform $X(\omega)$, e.g.,

$$\Phi(\omega)=\left|\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}x(t)e^{-j\omega t}dt\right|^2=\frac{1}{2\pi}X(\omega)X^*(\omega)$$

where $X(\omega)$ is the Fourier transform of $x(t)$, $X(\omega)=\int_{-\infty}^{\infty}x(t)e^{-j\omega t}dt$; $X^*(\omega)$ is the complex conjugate of $X(\omega)$.

The energy in $x(t)$ is $\int_{-\infty}^{\infty}x^2(t)dt$ or $\frac{1}{2\pi}\int_{-\infty}^{\infty}|X(\omega)|^2d\omega=\frac{1}{2\pi}\int_{-\infty}^{\infty}E(\omega)d\omega=\int_{-\infty}^{\infty}E(2\pi f)df$. The $E(2\pi f)$ is called the energy density spectrum.

Using the truncated $x(t)$ in $[-\frac{1}{2}T\ \ \frac{1}{2}T]$, the average power is given as

$$\lim_{T\to\infty}\frac{1}{T}\int_{-\frac{1}{2}T}^{\frac{1}{2}T}x_T^2(t)dt=\frac{1}{2\pi}\lim_{T\to\infty}\frac{1}{T}\int_{-\infty}^{\infty}|X_T(\omega)|^2d\omega=\frac{1}{2\pi}\int_{-\infty}^{\infty}S(\omega)d\omega=\int_{-\infty}^{\infty}S(2\pi f)df$$

where $S(\omega)$ is the power spectral density.

For a sequence $x[n]$, over an infinite number of elements, one has an energy spectral density as

$$\Phi(\omega)=\left|\frac{1}{\sqrt{2\pi}}\sum_{n=-\infty}^{\infty}x[n]e^{-j\omega n}\right|^2=\frac{1}{2\pi}X[k]X^*[k]$$

To obtain the ESD, we assume that the continuous and discrete Fourier transforms exist. This implies that $x(t)$ and $x[n]$ must be integrable (square-integrable) and summable (square-summable). To relax these requirements, the PSD is applied. The PSD is the Fourier transform of the autocorrelation function $R_{xx}(\tau)$, and

$$S(\omega)=\int_{-\infty}^{\infty}R_{xx}(\tau)e^{-j\omega\tau}d\tau=F(R_{xx}(\tau)),\ R_{xx}(\tau)=\lim_{T\to\infty}\frac{1}{T}\int_{-\frac{1}{2}T}^{\frac{1}{2}T}x(t)x(t+\tau)dt$$

The cross spectral density is the Fourier transform of the cross-covariance function. In particular, for *x* and *y* one has $P_{xy}(\omega)=\frac{1}{2\pi}\sum_{n=-\infty}^{\infty}R_{xy}e^{-j\omega n}$.

The ensemble average of the average periodogram when $T\to\infty$ is

$$S(\omega)=\lim_{T\to\infty}\frac{1}{T}E\left[\left|F\left((x_T(t)\right)\right|^2\right]=\lim_{T\to\infty}\frac{1}{T}E\left[\left|X_T(\omega)\right|^2\right],$$

where *E* is the expectation operator.

The PSD is a linear function of the auto-covariance and defines the variance. Furthermore, the power spectrum $G(x)$ is $G(x)=\int_{-\infty}^{x}S(x')dx'$.

Example 2. 1: Cross-Spectral Density

The proposed concept allows one to derive the cross-spectral density (CSD) function between two finite sequences as $S_{xy}(\omega)=\sum_{m=-\infty}^{\infty}R_{xy}(m)e^{-j\omega m}$.

While ESD and PSD are real-valued, the CSD is a complex functions. For finite $x[n]$ and $y[n]$ with different length, one may derive the confidence interval, and a coherent CSD is derived with a probability *p*. ■

Example 2.2: Robust Analysis of Sequences in the Frequency Domain

We apply the Fourier transform and examine the frequency components of a perfect nucleotide sequence $x[n]$, as well as $x[n]$ under uncertainties. The *symbolic quaternary alphabet* **A**={A,C,G,T} is mapped as **M**={0 1 2 3}. Figure 1.a reports a fragment of the studied $x[n]$. The magnitude of Fourier transform $|X[k]|$ indicates that there are four distinguished frequencies. Under uncertainties (gap, error, missing site, inconsistency, etc.), consider the sequence $x_\xi[n]$ which is documented in Figure 1.b. The comparison of resulting $|X[k]|$ for $x[n]$ and $x[n]$ indicates significant quantitative changes. However, the dominant frequencies can be identified, matching and similarity can be established, and, the sequence can be detected. Thus, data mining features are established.

The uncertainties (gap, error, missing site, inconsistency, etc.) are mapped by **U**. Let the uncertainties occur at 10, 25 and 40 sites. The gaps and inconsistencies can be mapped as **U**={–1 –2 …}. Let **U**={–1} with the resulting $x[n]$ as illustrated in Figure 1.c. The calculated

$|X[k]|$ is given in Figure 1.c. Thus, the frequency concept allows one to perform the analysis under very large uncertainties. Analysis of results and $|X[k]|$ allows us to conclude that the proposed concept ensures robustness, detection, observability, data mining and other features under large uncertainties. The qualitative and quantitative estimates correspond to a perfect sequence. The statistical methods may not provide relevant estimates and measures.

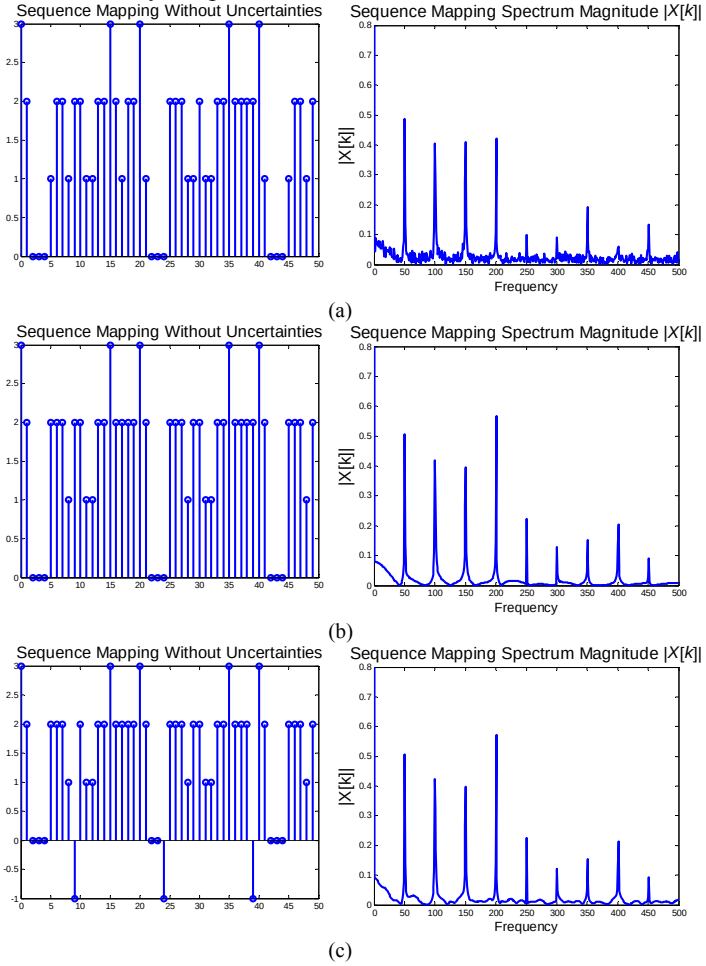


Figure 1. Nuceleotides mappings and Fourier transforms: (a) Perfect nucleotide sequence: Mapping $x[n]$ and $|X[k]|$; (b) $x_\xi[n]$ with errors and corresponding $|X[k]|$; (c) $x[n]$ with large uncertainties and resulting $|X[k]|$ ■

3. Applications and Results

The frequency-domain analysis was performed for complete *E.coli* and *S.typhimurium* genomes with 4,639,221 and 4,937,381 base pair strains in [3, 4]. An interactive toolbox is developed in MATLAB to accomplish a robust frequency-domain analysis. The sequences may not be complete, there can be missed sites, etc. The HIV and cancer genes are typical examples [5, 6]. It is virtually impossible to analyze patterns using statistical and "learning" methods. Furthermore, linear maps may not be found. In contrast, the reported concept is effectively applied providing meaningful results.

We perform the spectral analyses using ESD, PSD and other measures and estimates. Various parametric (autocorrelation, covariance, etc.), non-parametric (periodogram, Welch, etc.) and space methods are applied and utilized to obtain PSD. Figures 2 and 3 illustrate the PSDs for the nucleotide and amino acid sequences for HIV and cancer genes. The frequency analysis promises to solve a spectrum of problems such as: (1) Detect, identify and distinguish proteins and genes; (2) Examine and identify protein coding genes; (3) Potentially define structural and functional characteristics; (4) Analyze the data and perform data mining; (5) Identify patterns in gene sequences; (6) Enable classification; etc.

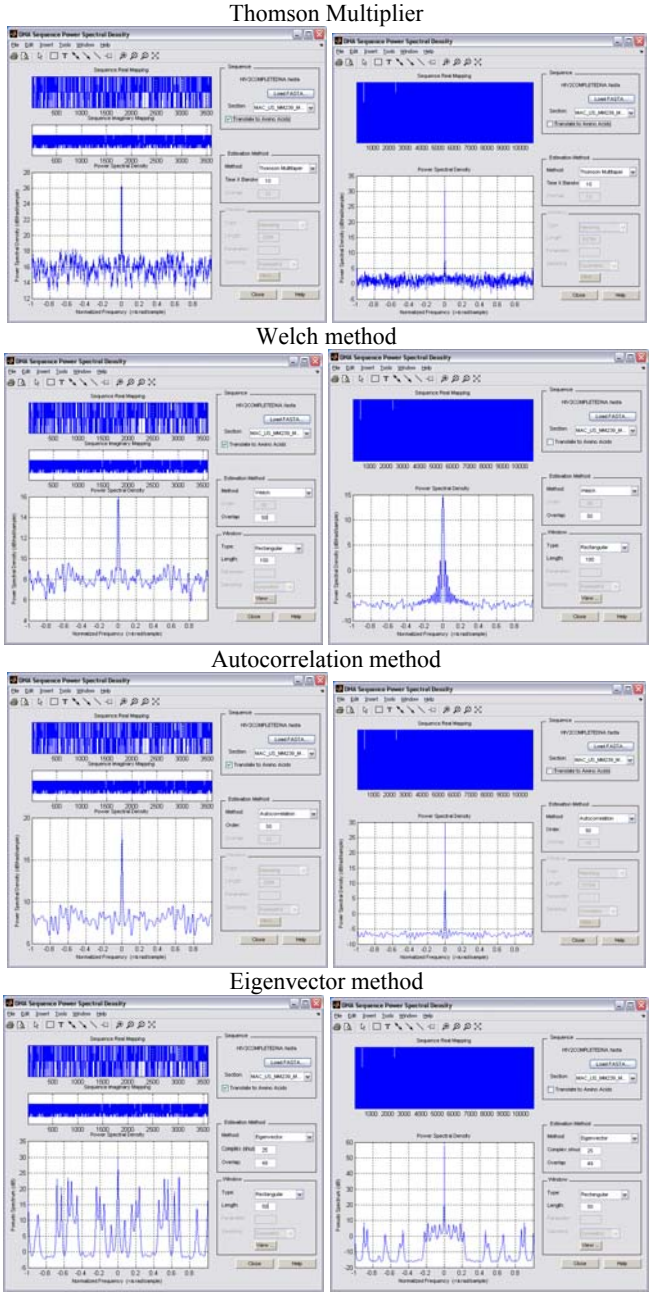
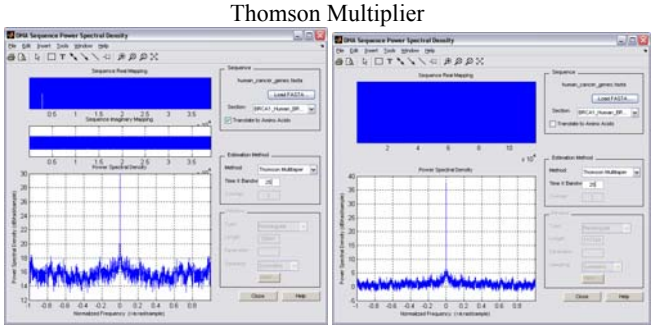
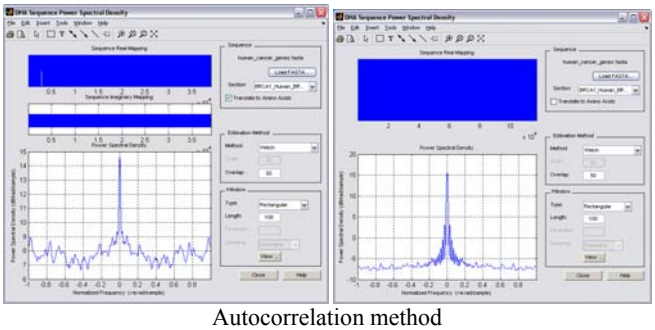


Figure 2. Power Spectral Density for the HIV sequence using amino acid and nucleotide sequences

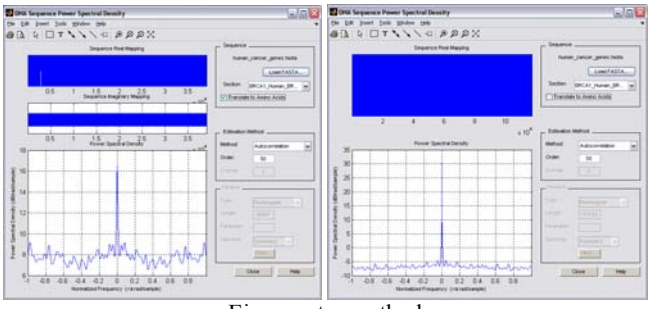
Thomson Multiplier



Welch method



Autocorrelation method



Eigenvector method

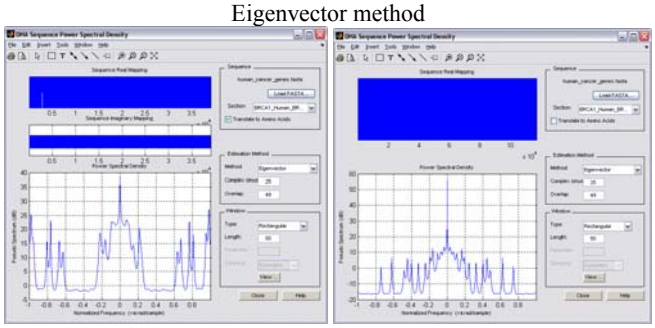


Figure 3. Power spectral density for the cancer sequence using amino acid and nucleotide sequence

One can examine spectral leakage, resolution, variance, modes, coherence and other spectral estimates and measures even over short or inadequate sequences. A data-intensive large-scale study of proteins may be quantitatively and qualitatively examined using the proposed approach. The protein structures and functionality can be potentially classified and identified. The proteomic analysis is more complex due to protein diversity and protein-protein interactions [1, 2]. While a genome does not evolve, a proteome differs from cell to cell and undergo changes (modifications, degradations, etc.) through various transitions, interactions and events with the genome and the environment. The number of proteins is much higher than genes. The increased complexity motivates the development of alternative approaches. The solution of this problem will affect the discovery of biomarkers, disease treatments, diagnostics, etc. For example, the genome and proteome information can be used to identify or implicate proteins associated with a disease. Specific and customized drugs can be designed to interfere, refine or inactivate the protein functionality. Drugs were found to target and inactivate the HIV-1 protease (an enzyme that cleaves a very large HIV protein into smaller functional proteins). The proposed concept promises to enable the homology and matching analyses, data mining, protein-protein evolutionary matching, profiling, classification and other tasks using sequenced and unsequenced proteins from genomes. There are needs for further studies, assessments and evaluation.

4. CONCLUSIONS

We proposed solutions to important problems in robust quantitative genome and proteome analyses. Our approach contributes to *bioinformatics* by developing a consistent fundamental concept. The spectral-centric analyses were performed. Cmplex sequences and patterns were robustly described and examined under uncertainties. These analyses promise one to enable: (i) Pattern recognition; (ii) Classification; (iii) Identification; (iv) Prototyping, etc. The proposed approach is useful due to: (1) Robust homology search and detection with high accuracy under uncertainties; (2) Accurate data-intensive analysis and evaluation; (3) Analysis of multiagent pathways for multi-genes; (4) Multifunctional analysis; (5) Computational efficiency and mathematical consistency; (6) Information extraction and information retrieval; (7) Large-scale capabilities using multiple databases; (8) Regressive and correlation analyses; etc. The proposed approach was found to be consistent, coherent, robust, compact and illustrative.

REFERENCES

1. W. P. Blackstock and M. P. Weir, "Proteomics: quantitative and physical mapping of cellular proteins", *Trends Biotechnol.* , vol 17, no. 3, pp. 121-127, 1999.
2. R. M. Twyman, *Principles of proteomics*, BIOS Scientific Publishers, New York, 2004.
3. S. E. Lyshevski, "Entropy-enhanced genome analysis in frequency domain," *Proc. NanoTech Conf.*, Boston, MA, vol. 2, pp.325-328, 2006.
4. S. E. Lyshevski and F. A. Krueger, "Robust entropy-enhanced frequency-domain genomic analysis under uncertainties," *Proc. IEEE Conf. Nanotech.*, Munich, Germany, pp. 556-558, 2004.
5. *Access to Complete Genomes and Proteomes*, European Bioinformatics Institute. http://www.ebi.ac.uk/proteome/ *FASTA/S Search* www.ebi.ac.uk/fasta33/proteomes.html
6. *HIV Databases*, Los Alamos National Laboratory www.hiv.lanl.gov